Veterinary Oncology

RESEARCH

Open Access

Physical activity predictors of cancer in Golden Retrievers: it's about frequency and intensity, not type

Dennis Ronzani[1] and Sarah E. Hooper[2*]

Abstract

Background Canine cancer is a leading cause of canine deaths, often resulting from complex interactions between germline-risk genetics, somatic mutations, and environmental exposures. To help identify major dietary, genetic, and environmental exposure risk factors for canine cancer, Morris Animal Foundation launched the Golden Retriever Lifetime Study, the first prospective longitudinal study in veterinary medicine. We hypothesized that responses from the physical activity section of the GRLS annual questionnaire could be used to develop a BiMM forest model that accurately classifies which Golden Retrievers develop cancer within the first seven years of the study. Furthermore, we expected that the most important predictors of cancer development would be the frequency and duration of the physical activity, with more rigorous activities—such as swimming—would be the most important predictors of cancer development.

Methods Activity and lifestyle questionnaire data for 3,044 purebred Golden Retrievers enrolled in the Golden Retriever Lifetime Study were obtained from Morris Animal Foundation. Two BiMM forest models were developed to predict the development of cancer: the "Years 0–7" model using consistently asked questions over the seven years, and the "Years 3–7" model, which incorporated additional questions about the pace and duration of physical activity starting in study year 3.

Results Of the enrolled dogs, 277 were diagnosed with cancer. The "Years 3–7" model achieved the best performance, with overall accuracy of 80.7%, a F1 score of 74.9% and a fair ROC AUC of 0.763. Key predictors of cancer development included year in study, frequency, pace, duration, and the frequency of warm and cold weather swimming. After Golden Retrievers were diagnosed with cancer, owners reported an 8–10% increase in exercise frequency and a 15.6% to 68.88% increase in cold weather swimming whereas warm weather swimming decreased by 2.0% to 13.9%. Similar declines in the pace and duration were also observed. The surface type where the exercise took place and the specific types of physical activity were lower in importance.

Conclusions Including pace and duration of the physical activity improved model performance, highlighting these predictors as high importance alongside the frequency of the physical activity. Future prospective studies should seek to determine specific physical activity guidelines for dogs, focusing on frequency, duration, and pace to potentially reduce cancer risk.

Keywords Canine cancer, Neoplasia, Machine learning, Prediction models, Exercise, Activity, Golden Retriever Lifetime Study, Golden Retriever, Environmental exposure, Canine cancer risk factors

*Correspondence:
Sarah E. Hooper
shooper@astate.edu
Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Neoplasia is a leading cause of canine deaths [1–5], with retrospective and survey-based research studies documenting that cancer accounts for 14.9% to 39% of all canine deaths in the examined populations [1, 4]. Determining the true incidence rates of spontaneous cancers remains challenging. This is due to factors such as variable access to veterinary care, inconsistent screening approaches due to owner financial constraints or clinician recommendations, and the absence of a national database for canine deaths [6]. Canine cancer is a complex, often multifactorial disease process [7, 8], and some risk factors vary based upon the specific type of neoplastic disease. For example, 50–70 percent of all neoplastic processes in intact female dogs are mammary tumors [7, 9]. Non-spayed females are at a greater risk of developing cancer compared to those spayed before their first heat cycle [10]. Schneider and colleagues reported mammary tumors occur in only 0.05% of females spayed prior to their first heat cycle, but this incidence rose dramatically to 8% and 26% if spaying was delayed until after the first or second heat cycle, respectively [10]. Ovarian estrogen and progesterone drive this increased risk by stimulating mammary duct and lobe proliferation and growth [11].

Furthermore, certain breeds are at a significantly increased risk of developing cancer, suggesting a breed-associated mode of inheritance for some cancers [7]. Golden Retrievers, for instance, show an increased incidence of hemangiosarcoma and lymphoma diagnoses [3, 5, 12]. Approximately 50 percent of Golden Retrievers deaths are attributed to cancer based on data from the Veterinary Medical Database (VMDB), a database containing abstracted medical record information from participating veterinary teaching hospitals [5]. Other retrospective studies assessing breed-related causes of death document similar mortality rates in Golden Retrievers [3, 13].

Due to their high reported cancer incidence and other factors such as breed popularity, researchers selected Golden Retrievers as the study breed for a Morris Animal Foundation (MAF) project designed to focus on four canine cancers: lymphoma, hemangiosarcoma, high-grade mast cell tumors and osteosarcoma [14]. In 2012, MAF launched the Golden Retriever Lifetime Study (GRLS), the first prospective longitudinal study in veterinary medicine [15]. Considering that canine cancer involves complex germline-risk genetics and somatic mutations driven partly by environmental exposure [7], GRLS aims to identify major dietary, genetic, and environmental risk factors for canine cancer and other diseases [15].

Longitudinal cohort studies like GRLS generate large amounts of data, or "big data", and offer researchers the ability to identify and relate disease diagnoses to specific lifestyle choices and environmental exposures [16]. With the advent of electronic medical records, fitness trackers, and other emerging technology, the generation of "big data" continues to grow the type of data contributing to clinical informatics—an emerging field where clinicians can make quicker and more accurate decisions about diagnosis, treatment, and prognosis of their patients using all types of patient data [17].

In clinical informatics and human oncology, relatively few studies have applied machine learning (ML) to analyze the "big data" generated from longitudinal cohort studies for predicting cancer risk, detecting cancer (including early detection), or identifying cancer recurrence [18]. Instead, the rapid expansion of ML algorithms has primarily focused on "big data" studies that assess multi-omics datasets (e.g. proteomic, genomics, transcriptomics, and metabolomics) and clinical data (e.g. serum chemistry) to discover modifiable risk factors, biomarkers, and disease prognostic indicators [19, 20].

Analyzing longitudinal cohort data, such as the MAF GRLS, presents challenges due to clustering and correlated observations, including serial correlation from repeated measures of each enrolled subject [21]. Applying classification ML methods to this type of data is particularly difficult, because many ML algorithms assume that each datapoint is independently sampled from a population [21], an assumption violated in longitudinal cohort studies with repeated sampling of the enrolled subjects. Classical statistical models also have limitations; some methodologies cannot be applied to longitudinal cohort datasets when the number of observations is smaller than the number of predictors [22].

Given these difficulties, researchers and veterinarians have largely underutilized the MAF GRLS owner questionnaire, which owners complete annually. Our study leverages a novel methodology developed by Speiser et al., which combines the generalized linear mixed model (GLMM), a classical statistical model, with the random forest ML algorithm. This binary mixed model (BiMM) forest, has been shown to effectively analyze clustered, high-dimensional data with interactions between predictors and nonlinear relationships between predictors and the outcome of interest [22]—specifically in our study, which Golden Retrievers will develop cancer.

This study aims to use the MAF GRLS annual owner questionnaire to determine whether the BiMM forest can build cancer prediction models that identify the key physical activity predictors of neoplastic diagnoses. Most research on canine physical activity research focuses on physiological responses, such as hematological changes, antioxidant defenses, and body condition

[23]. Few studies examine the relationship between physical activity and a disease, with most limited to areas such as canine osteoarthritis and obesity [24]. With no prior investigations into a link between canine physical activity and cancer development, we hypothesize that owner-reported responses from the physical activity section of the GRLS annual questionnaire can be used to develop a BiMM forest model that accurately classifies which enrolled Golden Retrievers would develop cancer within the first seven years of the study. Furthermore, we hypothesized that the most important predictors of cancer development will be the frequency and duration—amount—of physical activity with more rigorous types of physical activities such as swimming would be the most important predictors of cancer development.

Methods

Study population

We obtained data from 3,044 purebred Golden Retrievers enrolled in the MAF GRLS. Guy et al. [15] provides a detailed description of this observational cohort study [15]. In brief, privately owned dogs were eligible for enrollment if their pedigrees were known for at least two generations, lived in the contiguous United States, and were between six months and two years of age at the time of enrollment in the study. Owners and their veterinarians were required to commit to filling out an annual internet-based questionnaire, completing annual veterinary physical examinations, and collecting biological (e.g. blood and fecal) samples. MAF's Animal Welfare Advisory Board reviewed and approved the study. Informed consent was obtained from all owners and their veterinarians before enrolling each Golden Retriever in the study.

Raw data

All data was obtained from MAF's Data Commons Portal. At the time of this study, the portal included data from years zero through seven, with future years unavailable due to an embargo period. We analyzed data from the "Dog Demographics" (Supplemental Table 1) and "Activity and Lifestyle" (Supplemental Tables 2, 3, 4 and 5) sections of the Annual Owner Questionnaire Survey, and "Conditions Neoplasia" dataset which is data from the Annual Veterinarian Questionnaire, Additional Veterinarian Visit Questionnaire, Malignancy Related Questionnaire and Death and Necropsy Related Questionnaire. Changes to activity-related questions on the annual owner questionnaire resulted in three "Activity and Lifestyle" datasets: 1) Activity Overview—questions held consistent years zero through seven (Supplemental Table 6), 2) Activity Details Through SY2—data from questions addressed years zero through two (Supplemental Table 7), and 3) Activity Details SY3 Beyond—data from questionnaire addressed questions beginning from year three (Supplemental Table 8). We conducted all data preparation, preprocessing, and analysis using Microsoft Excel 365 and R (version 4.3.2) [25] within the integrated development environment RStudio (version 2023.6.0.421) [26].

Data preparation

Prior to merging the following three activity datasets 1) Activity Overview, 2) Activity Details Through SY2, and 3) Activity Details SY3 Beyond, we modified the structures to be uniform. Specifically, we converted the reported binary fetch and binary walking activity variables in the first two years into the format reported for years three through seven as illustrated in Table 1. Questions where the owner response rate was lower than 50% were not included in the analysis (Table 1).

Upon merging all datasets, we removed duplicate rows and manually reviewed all 1,428 free text entries in the "other specify" column. Activities that referred to one of the standard activity options were revised to the appropriate format/column. For activities labeled as swimming, we entered the activity under the "swim" variable, and recorded frequency and location when available. We grouped the 138 unique entries for other activities under a single "other activities" variable. To ensure consistency, we standardized total time entries for each row by adjusting all total time entries to reflect only the reporting year. We calculated each dog's age and additional details for the data preparation are outlined in the Supplemental data.

Data preprocessing

The Binary Mixed Model (BiMM) random forest models created for this study are only able to use numerical data, therefore we converted all non-numerical data to a numerical format using the processes described in this section. We performed ordinal encoding of ranked factors using the R package dplyr (version 1.1.4) [27] and one-hot encoded the activity variable using the R packages data.table (version 1.14.10) [28] and mltools (version 0.3.5) [29]. Data from years zero through seven was combined into one dataset (Data Years 0–7), and selected variables are listed in Table 1 and Supplemental Table 6 and 7. Data from years three through seven was combined into a second dataset (Data Years 3–7), with variables listed in Supplemental Table 8.

Missing values in Data Years 0–7 dataset and Data Years 3–7 dataset were imputed using multivariate imputation by chained equations (MICE) with the random forest algorithm within the mice package (version 3.16.0) [30]. We performed five imputations with 30 iterations

Table 1 The reported raw variables in the dataset "Activity Details Through SY2" with details on how they were converted to be in the same format as the "Activity Details SY3 Beyond" variables

Activity details through SY2 variable	Conversion	Activity details SY3 beyond variable
subject_id	Did not change	subject_id
year_in_study	Did not change	year_in_study
record_date	Did not change	record_date
activity_level	Did not change	activity_level
walk_frequency	Changed to frequency	frequency
walk_duration	Changed to duration	duration
walk_pace	Average on lead/leash walk pace	pace
walk_reason_bathroom	Binary variable changed to "On leash (obedience training, walking, or running)" and reported under the activity variable	activity
walk_reason_general_enjoyment	Binary variable changed to "On leash (obedience training, walking, or running)" and reported under the activity variable	activity
walk_reason_training	Binary variable changed to "On leash (obedience training, walking, or running)" and reported under the activity variable	activity
walk_reason_other_specify	Free text variable changed to "On leash (obedience training, walking, or running)" or "Off leash (obedience training, walking, or running)"	activity
walk_without_leash	Binary variable changed to "Off leash (obedience training, walking, or running)" and reported under the activity variable	activity
aerobic_duration	Eliminated due to few owner responses	-
aerobic_pace	Eliminated due to few owner responses	-
aerobic_frequency	Eliminated due to few owner responses	-
fetch_games_frequency	Character value reporting the frequency of fetch changed to "fetch" and reported under the activity variable and under the frequency variable	activity and frequency

for each dataset. To balance each dataset, we applied the synthetic minority over-sampling technique for nominal and continuous variables (SMOTE-NC) using the RSBID package (version 0.0.2.0) [31]. The final k-values for SMOTE-NC were based on model performance, using k = 7 for Data Years 3–7 and k = 3 for Data Years 0–7. Continuous variables were scaled between 0 and 1 using the caret package (version 6.0.94) [32]. This ensured all variables were on the same scale.

Each imputed dataset was divided into training and testing sets. The training dataset contained 50 percent of the Golden Retrievers with cancer and 50 percent without cancer. To prevent overlap, no subject patient included in the training set appeared in the testing set, ensuring data independence through a cluster approach [33].

Model construction

BiMM random forest models were constructed using the R code obtained from Speiser et al. 2019 [22]. We employed the H3 method with an error tolerance of 0.01 and one iteration, as recommended [22]. BiMM models were run on each of the five imputations of Data Years 0–7 dataset (Supplemental Table 6) and Data Years 3–7 dataset (Supplemental Table 7 and 8) and pooled the results into a single confusion matrix. This matrix was used to calculate the overall accuracy, sensitivity/recall, specificity, precision, F1 and the area under the receiver operating characteristic curve ROC Curve (AUC) along with the 95% confidence intervals [34].

Variable exploration by year

We calculated the mean (with standard deviation) and the median (with range) for the top five most important exercise variables—frequency, duration, pace, swim frequency in warm weather, and swim frequency in cold weather—for dogs with and without a neoplastic diagnosis by study year. These calculations were performed using the R package *dplyr* (version 1.1.4) [27].

To determine the average lifetime percent change for each variable, we first calculated the average values reported in the year(s) prior to the diagnosis and the year(s) post-diagnosis. We then subtracted the pre-diagnosis average from the post-diagnosis average and divided this difference by the sum of the two averages. This calculation provided the percent change in exercise behavior over the dog's lifetime following the cancer diagnosis.

Results

Demographics

A total of 3,044 Golden Retrievers were enrolled in the study with 1 cancer diagnosis the first year of the study.

At baseline (year 0), 219 were intact females, 431 were intact males, 1,109 were neutered males, and 1,285 were spayed females. Cancer diagnosis sequentially increased over the first seven years of the study with 277 enrolled dogs being diagnosed with cancer (Table 2). The median age of a cancer diagnosis was 6.1 years (Table 2).

Model performance

We constructed two BiMM random forest models. The model using variables consistent over Years 3–7 achieved the best performance, with overall accuracy of 80.7%, a F1 score of 74.9% and a fair ROC AUC of 0.763. The model using variables consistent over years 0–7 performed poorly with an overall accuracy of 68.2%, a F1 score of 56.4%, and a ROC AUC of 0.622. We did not further develop or explore the Years 0–7 model, due to its poor performance (Table 3). Table 3 summarizes the calculated performance metrics for both models.

Years 3–7 Most important predictors for development of cancer

The larger the mean decrease Gini value, the more important the variable. Of the top 10 most important predictors for classifying if a Golden Retriever had cancer or did not have cancer, 4 were directly related to the frequency and duration of aerobic physical activity. The top 7 most important predictors for developing cancer in descending order were year in study, frequency, pace, duration, the frequency of warm and cold weather swimming, and the overall activity level reported by the owner (Fig. 1). The surface where the activities took place and the specific activity types ranked lower in importance. Figure 1 displays all variables included in the final Year 3–7 model.

Exercise frequency

Exercise frequency was recorded over all 7 years and increased over time, with dogs exercising weekly to less than once per week at entry into the study progressing to daily exercise by year seven. Over years three through seven, the exercise frequency for all activities (excluding swimming) showed minimal variation between Golden Retrievers with a cancer diagnosis and those without cancer (Fig. 2A, Supplemental Table 9). However, the average lifetime changes in exercise frequency starting the year after diagnosis notably increased by approximately 8–10 percent for dogs diagnosed between years one and five but decreased for those diagnosed in year six (Fig. 2B, Supplemental Table 9).

Exercise pace

Exercise pace was recorded only from year three and onwards. Among years three through seven, the exercise pace for all activities was classified by the owners as brisk/brisk walk and the average varied minimally between Golden Retrievers with a cancer diagnosis and those without cancer (Fig. 3A, Supplemental Table 11). The average lifetime changes in exercise pace showed a

Table 2 The number of enrolled golden retrievers during the first seven years of the MAF GRLS

Year of study	Number of enrolled dogs	Age range of enrolled dogs	Number of cancer diagnosis	Mean age of diagnosis (SD)	Median age of diagnosis (Range)
0	3,044	0.4–3.3	1	1.1	1.1
1	2,706	1.4–4.3	14	2.6 ± 0.6	2.8 (1.5—3.5)
2	2,604	2.4–4.0	14	3.5 ± 0.6	3.8 (2.6—4.0)
3	2,494	3.4–5.0	29	4.5 ± 0.6	4.5 (3.5—5.5)
4	2,379	4.4–6.0	41	5.3 ± 0.5	5.3 (4.6—6.2)
5	2,187	5.4–7.0	59	6.5 ± 0.6	6.4 (5.5—7.8)
6	1,992	6.4–8.0	49	7.3 ± 0.6	7.4 (6.6—8.3)
7	1,943	7.4–9.0	70	8.4 ± 0.4	8.3 (7.6—9.1)
All Years	Up to 3,044	0.4–9.0	277	6.0 ± 1.7	6.1 (1.1—9.1)

The age range of all enrolled dogs reported alongside the mean with the standard deviation (SD) and median age with the range of enrolled Golden Retriever subjects when cancer was diagnosed

Table 3 Performance metrics for Years 0–7 and Years 3–7 models with 95% confidence intervals

Model	Accuracy	Sensitivity/recall	Specificity	Precision	F1	AUC
Years 0–7	68.2% (68.1%, 68.4%)	56.4% (56.0%, 56.8%)	74.9% (74.7%, 75.3%)	56.4% (56.2%, 56.6%)	56.4% (56.2%, 56.6%)	0.622
Years 3–7	80.7% (80.4%, 80.8%)	84.9% (84.8%, 85.1%)	78.6% (77.8%, 78.9%)	67.3% (67.2%, 67.4%)	75.0% (74.9%, 75.2%)	0.763

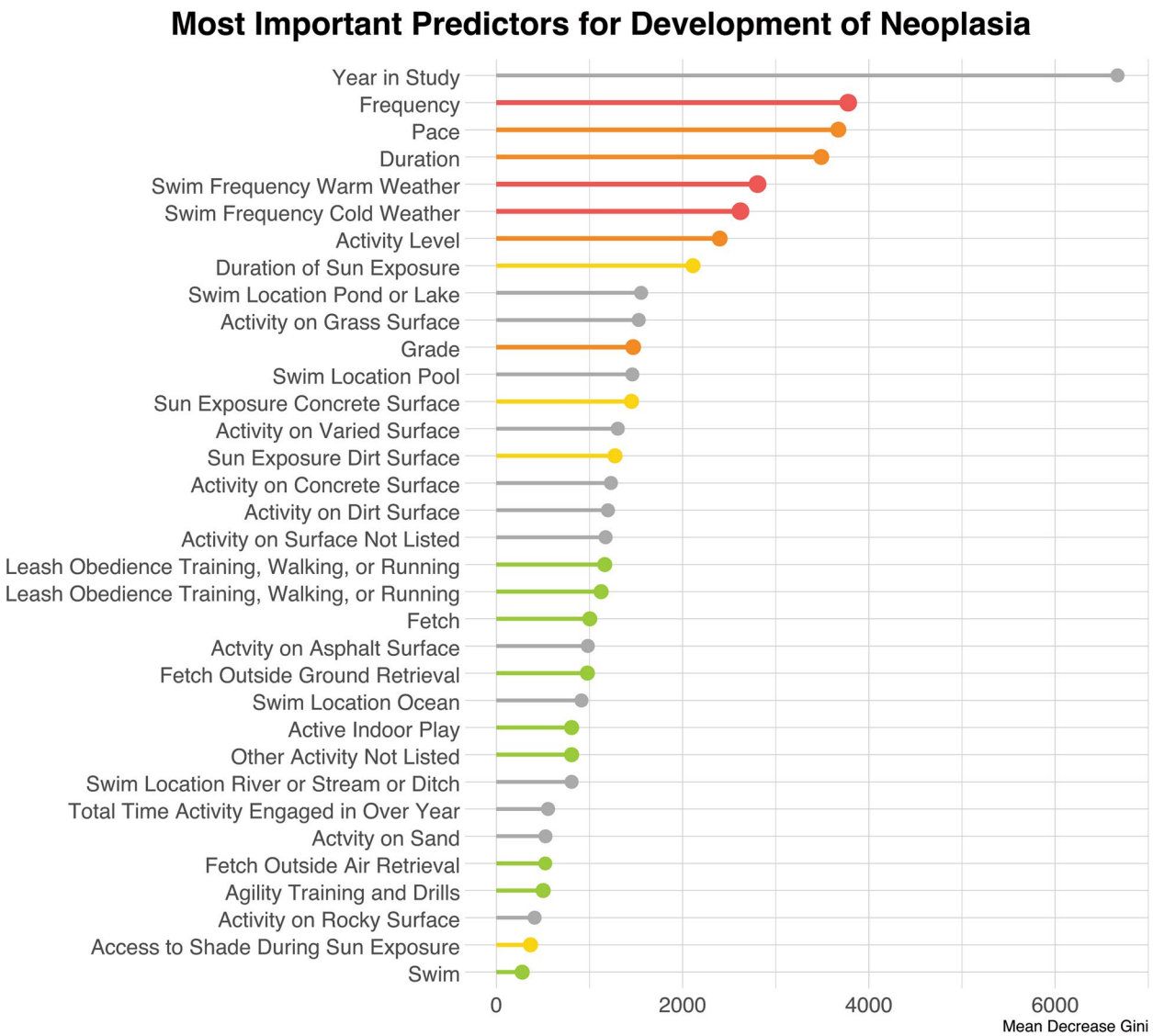

Fig. 1 Predictors included in the Year 3–7 Model, ranked by their mean Gini importance. Larger Gini values indicate higher relative importance of the predictors to the model's performance. The light red variables relate to the frequency of a physical activity. The orange variables describe the pace, duration, and grade of the physical activity. The yellow variables relate to sun exposure. The green variables describe the type of physical activity, and all other variables describing the location of the activity are in gray

slight increase for Golden Retrievers diagnosed in year 4 (Fig. 3B). For those diagnosed in years five and six, the average exercise pace decreased by 5.7 percent and 2.1 percent, respectively (Fig. 3B, Supplemental Table 12).

Exercise duration

Exercise duration was recorded starting in year three. During years three through seven, the mean and median exercise duration for all activities ranged between 10–60 min. While the duration showed minimal variation between Golden Retrievers with a cancer diagnosis and those without cancer (Fig. 4A, Supplemental Table 13), the average lifetime changes in exercise duration consistently decreased overall, with dogs diagnosed in year four exhibiting the greatest decline—a 7.8 percent reduction in exercise duration post-diagnosis (Fig. 4B, Supplemental Table 12).

Swim frequency

All dogs swam more frequently in warm weather, with frequencies ranging from rarely to daily, compared to cold weather, where frequencies ranged from never to monthly (Fig. 5A, Supplemental Tables 14 and 15). In

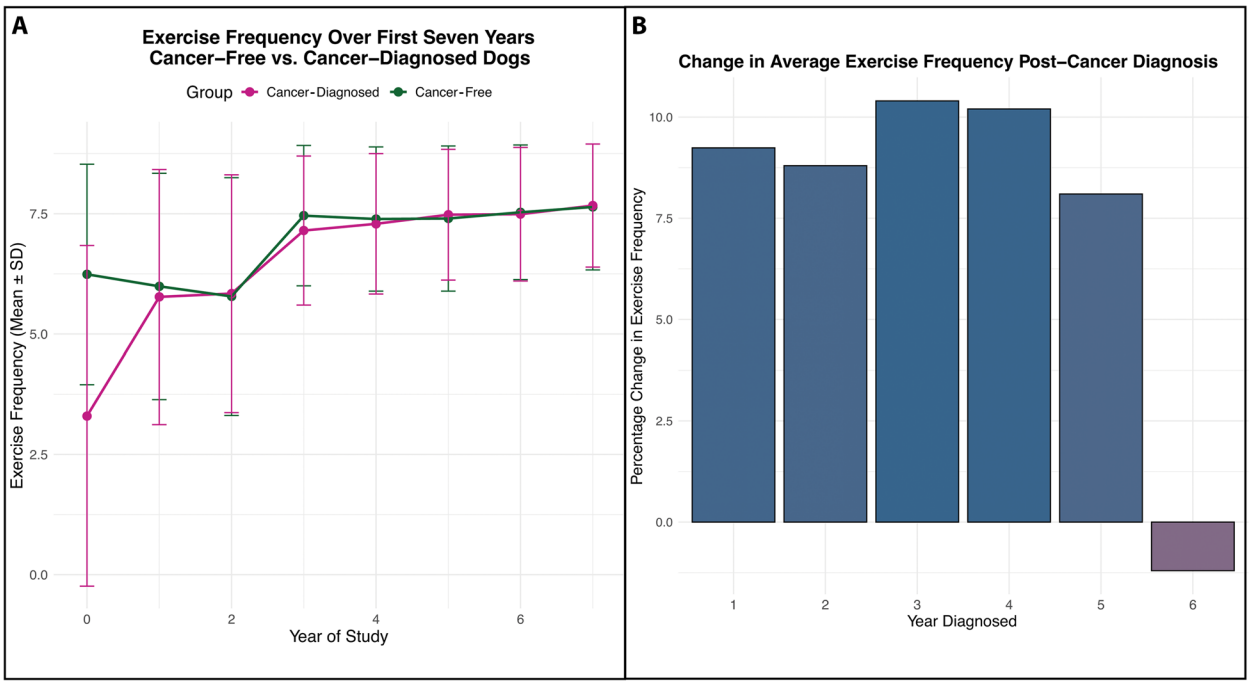

Fig. 2 **A** The frequency of exercise reported by owners in the survey is presented for Golden Retrievers without a cancer diagnosis and those with a cancer diagnosis. Golden Retrievers increased from the baseline of walking less than once/week to walking weekly in years 1 and 2 to more frequently beginning in year 3. Supplement Table 8 provides additional details on the ordinal encoding used to convert the owner categorical responses to numerical values. **B** Changes in average exercise frequency for Golden Retrievers diagnosed with cancer, expressed as percentage change from pre-diagnosis levels by study year of diagnosis

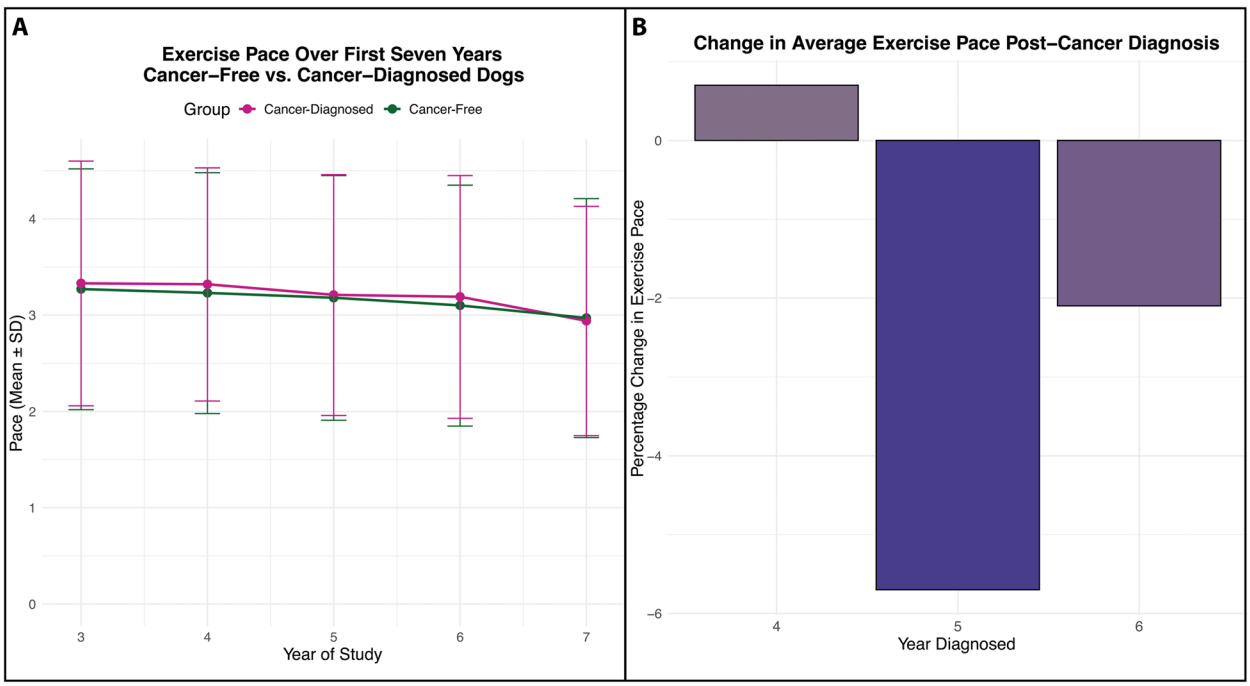

Fig. 3 **A** The pace of exercise reported by owners in the survey is presented for Golden Retrievers without a cancer diagnosis and those with a cancer diagnosis. **B** Changes in average exercise pace for Golden Retrievers diagnosed with cancer, expressed as percentage change from pre-diagnosis levels by study year of diagnosis

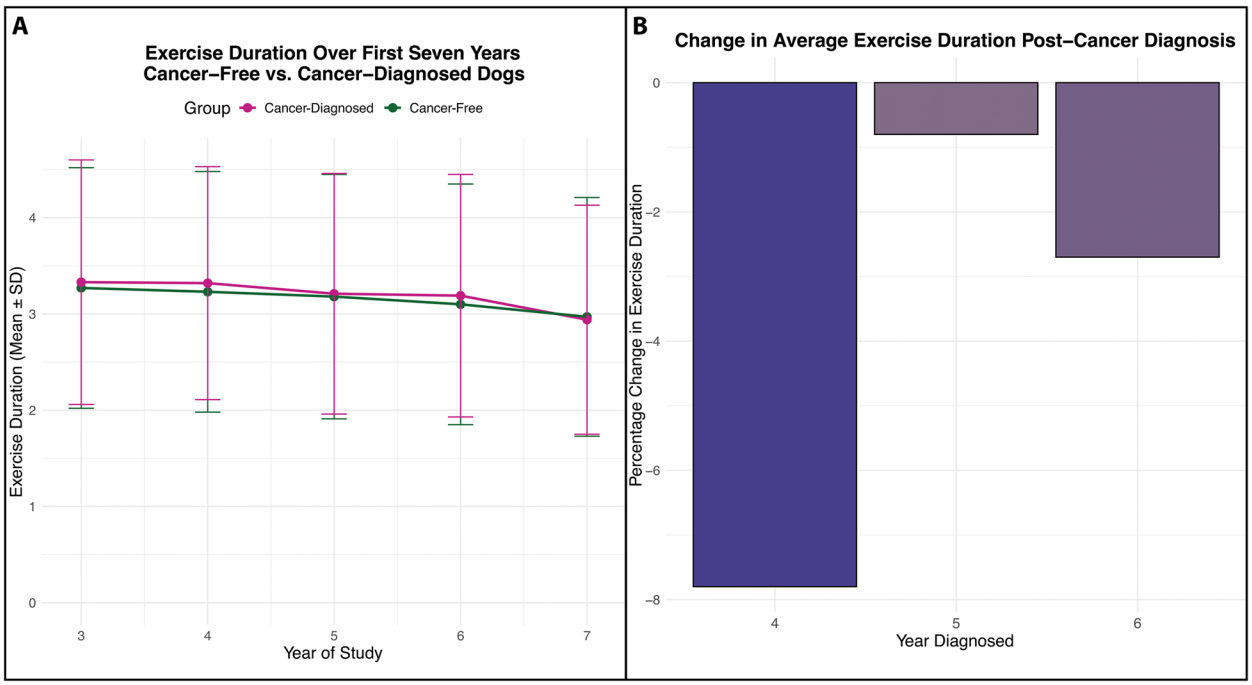

Fig. 4 **A** The duration of exercise reported by owners in the survey for Golden Retrievers without a cancer diagnosis and those with a cancer diagnosis. **B** Changes in average exercise duration for Golden Retrievers diagnosed with cancer, expressed as percentage change from pre-diagnosis levels by study year of diagnosis

warm weather, Golden Retrievers diagnosed with cancer swam daily to monthly, while those without cancer swam rarely to monthly. Owners of dogs diagnosed with cancer during years three through seven reported a slightly higher frequency of swimming in warm weather (Fig. 5A). Over years one through seven, owners reported a slightly higher average swimming frequency in cold weather compared to those without a cancer diagnosis (Fig. 5A). Dogs diagnosed with cancer in year one and years three through seven swam less in warm weather but notably swam 15.6 percent to 68.88 percent more frequently in cold water during years one through six (Fig. 15B).

Activity level

Over the first seven years of the study, most dogs were reported as having little activity to moderate activity levels. Owners of Golden Retrievers without a cancer diagnosis reported the highest activity levels at study year zero (baseline), with many owners reporting the classification of very active activity levels. Activity levels gradually declined over the subsequent years (Fig. 6A, Supplemental Table 17). The activity levels of Golden Retrievers diagnosed with cancer did not show as consistent of a pattern of decline (Fig. 6A, Supplemental Table 17). However, once diagnosed with cancer, their lifetime activity level declined as much as 15.5%.

Sun exposure duration

At baseline and during the first year of the study, Golden Retrievers without cancer were exposed to the sun for three to eight hours per day. In subsequent years, their sun exposure decreased to less than three hours daily (Fig. 7A, Supplemental Table 18). At baseline and years two through seven, dogs diagnosed with cancer received less than three hours of sun exposure (Fig. 7A, Supplemental Table 18). The average sun exposure duration was consistently slightly lower for dogs diagnosed with cancer. The change in average sun exposure duration was highly inconsistent in dogs with cancer. Dogs diagnosed with cancer in year one and year six had notable decreases in sun exposure duration, while dogs diagnosed in year two through five had minimal changes (Fig. 7B).

Discussion

Canine cancers are one of the most devastating diseases, affecting one in four dogs [35, 36], and are a leading cause of death in dogs over 10 years [1, 2]. While most epidemiologic studies on canine cancer are retrospective, the MAF GLRS study offers a unique opportunity as a prospective cohort study [37]. Our study is the first to analyze owner-reported physical activity questionnaire data from all enrolled dogs in the MAF GRLS. Additionally, this is the first report to apply ML prediction models to explore potential environmental factors influencing cancer risks in Golden Retrievers.

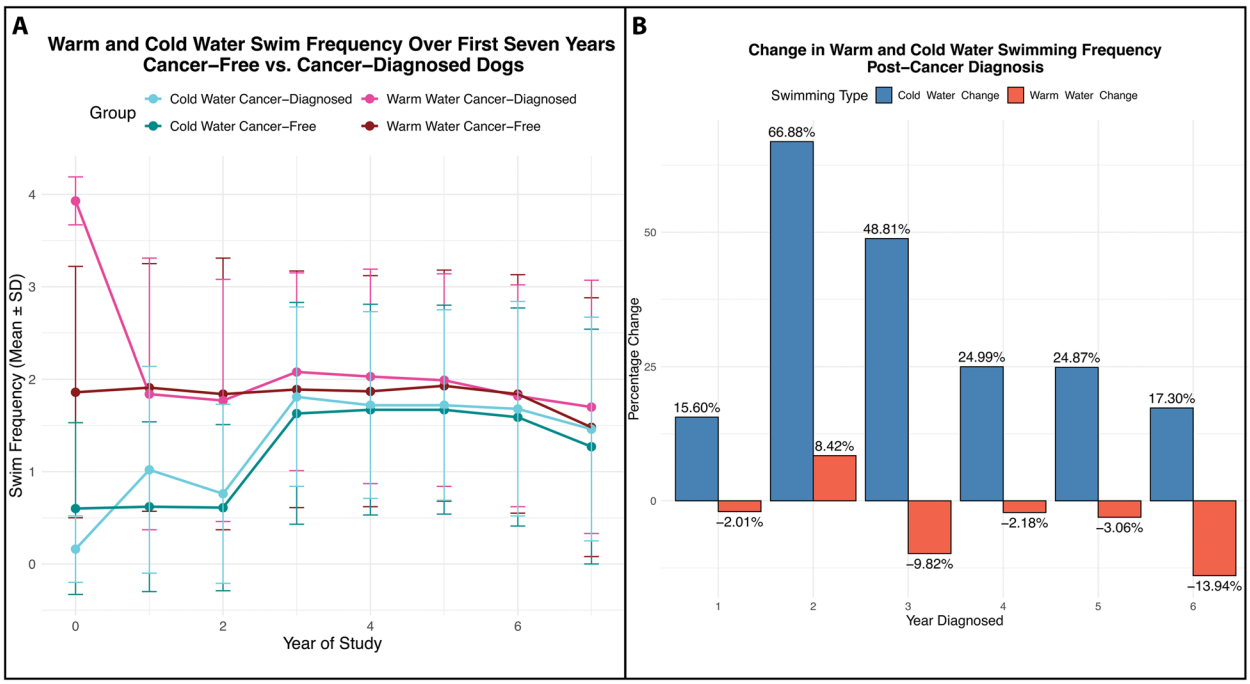

Fig. 5 **A** The swimming frequency in cold weather (blue) and warm weather (pink) reported by owners in the survey for Golden Retrievers without a cancer diagnosis and those with a cancer diagnosis. **B** Changes in average swimming frequency in cold (blue) and warm (red) for Golden Retrievers diagnosed with cancer, expressed as percentage change from pre-diagnosis levels by study year of diagnosis

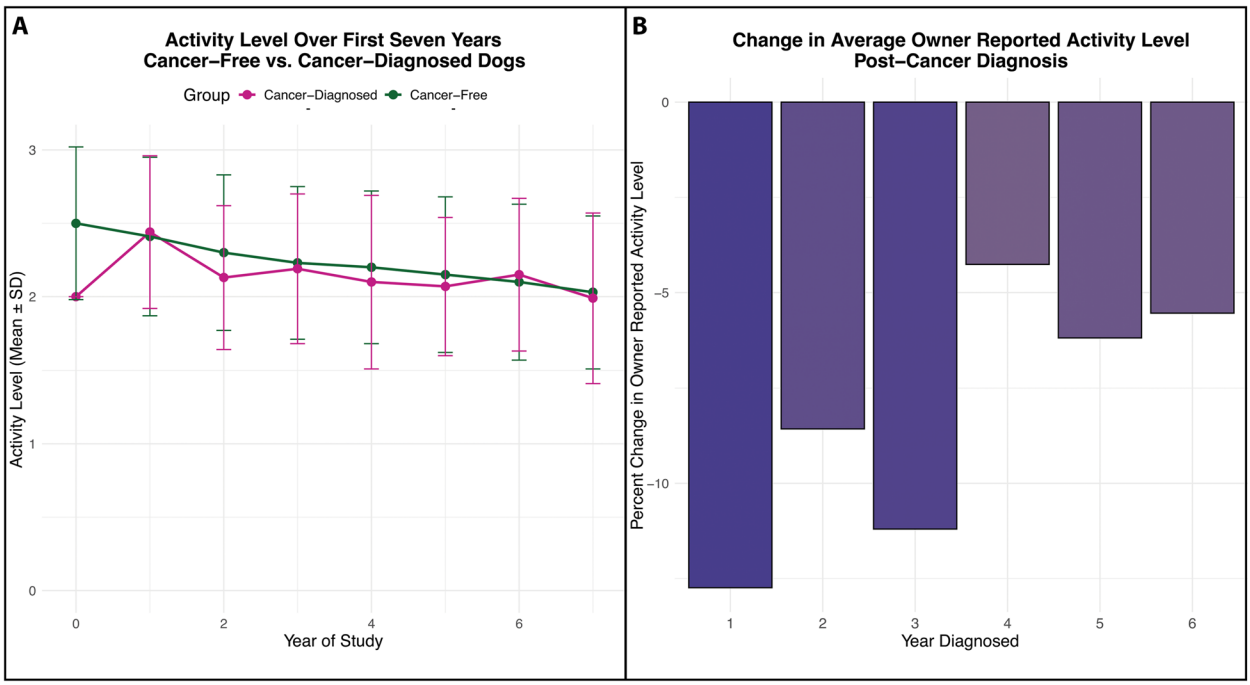

Fig. 6 **A** The activity level reported by owners in the survey for Golden Retrievers without a cancer diagnosis and those with a cancer diagnosis shows a decline over the course of the first 7 years of the study. **B** Decline in the activity levels for Golden Retrievers diagnosed with cancer, expressed as percentage change from pre-diagnosis levels by study year of diagnosis

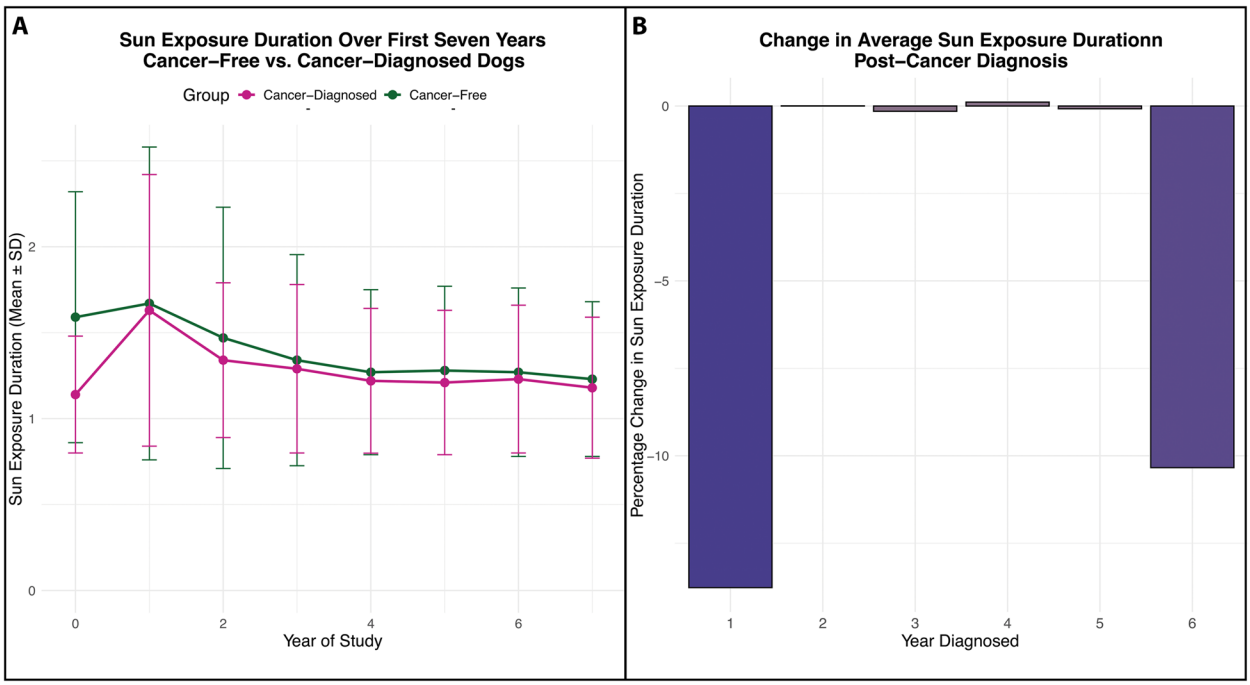

Fig. 7 **A** The sun exposure duration reported by owners in the survey is presented for Golden Retrievers without a cancer diagnosis and those with a cancer diagnosis. **B** The inconsistent changes in average sun duration exposure for Golden Retrievers diagnosed with cancer, expressed as percentage change from pre-diagnosis levels by study year of diagnosis

Moderate-to-vigorous physical activity has been strongly associated with a reduced risk of several cancers, including colon, intestinal, kidney, liver, and mammary cancers, even in individuals with a genetic predisposition in humans [38–40]. For example, in human medicine, moderate physical activity has been shown to delay cancer onset in women who are genetically predisposed to breast cancer [38]. Despite robust evidence in humans, the relationship between exercise and cancer development in dogs has never been explored. To address this knowledge gap, we investigated whether cancer development in Golden Retrievers could be predicted by physical activity patterns, the surfaces on which activities occurred, and the amount of daily sun exposure—all data within the MAF GRLS physical activity datasets.

The BiMM random forest model, using data available for years three through seven, achieved an overall classification accuracy of approximately 81% and a sensitivity of approximately 85%. This performance is good, considering that of the 3,044, only 248 were diagnosed with cancer between years three and seven. This performance aligns with the performance of classification ML algorithms applied to human cohort studies [22, 41, 42].

We selected the BiMM forest due to the statistical limitations of employing generalized linear mixed models (GLMMs) to clustered and longitudinal binary outcomes [22, 33]. BiMM forests effectively handle the interactions among variables and non-linear variable relationships with the outcomes of our study [22, 42], relating to diagnosis of cancer or no cancer diagnosis. Unlike the traditional random forest algorithm, which requires summarizing the predictors (e.g. averaging over all the time points) or selecting a single time point, the BiMM forest was designed to account for repeated measures or clustered patient data [22, 33].

Our BiMM forest models enabled us to assess the relative importance of activity predictors to answer our hypothesis. Study results confirmed that physical activity frequency, duration, pace (intensity) and swimming frequency in both warm and cold weather emerged as top predictors (Fig. 1) while surprisingly, the different types of physical activity were shown to be of relatively low importance (Fig. 1).

While no established physical activity recommendations or guidelines exist for dogs, there are several well-established guidelines for humans. For instance, the American Cancer Society recommends that adults "engage in 150 to 300 min of moderate-intensity physical activity per week, or 50 to 150 min of vigorous-intensity physical activity…achieving or exceeding the upper limit of 300 min is optimal" [39]. According to the United States Department of Health and Human Services, only 1 in 4 adults and 1 in 5 adolescents [43] meet the physical activity guidelines [39, 40]. Most GRLS owners reported

walking their dogs once to a few times a week at a brisk pace, which qualifies as moderate-intensity activity according to the Physical Activity Guidelines for Americans [40]. With the average physical activity duration of 31–60 min and a median of 10–30 min, many enrolled Golden Retrievers appear to engage in less activity than the recommended 150 to 300 min per week for humans. This observation mirrors the finding that 75 percent of the US population does not meet the physical activity guidelines [43].

Interestingly, Golden Retrievers diagnosed with cancer were less frequently physically active compared to the study average. This finding is obscured in large part because owners increased the frequency of exercise after their dog's cancer diagnosis, leading to an average post-diagnosis increase in the physical activity frequency by about 10% (Fig. 2A and B). Similar increases are observed in human patients with approximately three out of four human cancer survivors meeting the physical activity guidelines [44]. This post-diagnosis change in exercise could reflect recommendations from their veterinarian, or alternatively, owners could find information on websites and blogs emphasizing the benefits of physical activity for canine cancer. Few resources, apart from the Animal Cancer Foundation, acknowledge the lack of evidence for the effect of exercise on pets [45].

Starting in year three, MAF added questions about the duration and pace (intensity) of the physical activity to the annual owner questionnaire. Both predictors were identified as highly important (Fig. 1), likely explaining why the model using data from years 0 through 7 performed worse than the model using years 3 through 7. Our findings on frequency, intensity (equivalent to pace on the MAF GRLS questionnaire), and duration of the physical activity align with the few clinical trials that suggest these factors have measurable effects on cancer biomarkers.

Several case–control studies have shown that the duration of physical activity impacts the risk of breast cancer development. Women who reported exercising 2–3 h per week had an average 9% reduction in breast cancer risk, while those who reported exercising 6.5 h per week had a 30% reduction in risk [46]. However, these studies did not account for the specific duration of individual exercise sessions. Evidence from animal studies suggests that the duration of individual sessions may have less impact than the overall quantity of exercise. For instance, rats performing treadmill exercise at moderate-intensity levels demonstrated a lower incidence of mammary cancer and a reduction in cancer multiplicity, with no significant difference between groups running for 20 min or 40 min per day [47].

The Golden Retrievers in the GRLS participated in physical activity lasting 10–60 min on average and was conducted at a moderate-intensity level. Moderate-intensity physical activity is optimal for cancer prevention in other species and has been shown to inhibit cancer cell proliferation and induced apoptosis [48, 49]. While high-intensity physical activity can also be protective, strenuous physical activity may exacerbate certain cancers, such as breast cancer, by promoting the growth of cancerous cells in animal models [48, 50]. This suggests that the Golden Retrievers typically engaged in appropriately intense physical activity to support cancer prevention per accepted human guidelines. However, post-diagnosis, the observed decrease in intensity and duration (Figs. 3B and 4B) may have negated the potential benefits of increased exercise frequency.

The most notable increase in physical activity frequency was in cold weather swimming, with post-diagnosis of cancer, the Golden Retrievers swimming frequency increased by 15.60% to 66.88%. In humans, physical activity, including swimming, is well-established to reduce the severity of chemotherapy and radiotherapy side effects while concurrently improving survival [48]. The growing popularity of aquatic therapy, such as underwater treadmills [51], warrants further investigation. Future studies should consider surveying veterinary oncologists to determine whether aquatic therapy is commonly recommended as part of their patient's cancer treatment and rehabilitation.

The most common swimming locations reported were pools, ponds and lakes. According to the United States (US) Environmental Protection Agency (EPA) most recent *National Water Quality Inventory Report* to Congress, over 13 million acres of lakes and ponds (excluding the Great Lakes) are considered for activities such as swimming due to pollution [52]. Unfortunately, dogs can also be exposed to pollutants in pools. While researchers suggest the benefits of pool swimming outweigh the risks [53], more research is needed to determine if breeds predisposed to cancer should limit pool swimming. Epidemiologic studies in humans have linked disinfection-by-products, such as trihalomethanes, to increased cancer risk [54, 55]. Similarly, a study found that dogs with urothelial cell carcinoma were more likely to have swam in pools compared to dogs without cancer [56].

The average age of cancer diagnosis in dogs is 8.5 years, with a median of 8.8 years [57]. Due to an embargo period, this study assessed only the first seven years of the MAF GRLS data, meaning many dogs were yet to be diagnosed with cancer. With fewer than 10 percent of the study population diagnosed with cancer and a range of cancer types and grades represented, there were insufficient

cases to develop a ML model for any specific cancer type. Consequently, a limitation of this study is the grouping of all cancer types and grades together. BiMM forest model performance is likely to improve as more enrolled Golden Retrievers are diagnosed, enabling the creation of models specific to individual cancer types. After the data embargo expires, future studies should reassess these findings and the use of the BiMM forest model to determine whether the key predictors change and if model performance improves with additional years of data.

Another limitation of this study was the lack of body condition score (BCS) data for many of the enrolled dogs. Approximately 41 percent of the physical exam records did not report a BCS. Furthermore, BCS was only recorded for years three through seven, as the Hill's Body Fat Index was used during years zero through two. Additionally, many records lacked body fat index scores, and no published methods are available to convert body fat index scores to BCS, which further complicates the inclusion of this variable. Golden Retrievers are overrepresented in studies as being overweight [58], and low levels of physical activity, particularly when limited to walking, are associated with an increased risk of obesity [59]. Since obesity has been linked to the development of specific types of canine cancer [60], it would be worthwhile to educate the veterinarians participating in the study about the importance of complete physical exam records.

Conclusions

Our findings highlight that the frequency, pace, and duration of exercise, combined with exercise-related environmental factors over time, are strong predictors of cancer development in Golden Retrievers, suggesting that cumulative lifestyle exposures over time may play a critical role in cancer risk assessment.

These results underscore the importance of considering social determinants of health in future studies, as owner behaviors, geographic location, and access to certain environmental features may shape a Golden Retriever's activity profile. Additionally, future research should focus on identifying optimal activity levels and potential environmental exposure risks. By identifying integrating these findings, veterinarians can better guide pet owners on exercise recommendations and environmental modifications that may help reduce cancer risk in Golden Retrievers.

Abbreviations

BCS	Body condition score
EPA	Environmental Protection Agency
MAF	Morris Animal Foundation
ML	Machine learning
GRLS	Golden Retriever Lifetime Study
US	United States

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s44356-025-00024-5.

Supplementary Material 1.

Acknowledgements

We would like to thank Morris Animal Foundation, its staff members and all participants in the Golden Retriever Lifetime Study, including the dog owners, their golden retrievers and the Study veterinarians who made this work possible.

Authors' contributions

DR: Conceptualization, data curation, formal analysis, funding acquisition, investigation, writing – review and editing. SH: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, visualization, writing – original draft preparation, writing – review and editing. All authors read and approved the final manuscript.

Funding

The Golden Retriever Lifetime Study and this manuscript were made possible through financial support provided by the Morris Family Foundation, the Mark & Bette Morris Family Foundation, VCA, the V Foundation, Blue Buffalo Company, Petco Love, Zoetis, Antech Inc., Elanco, the Purina Institute, Orvis, the Golden Retriever Foundation, the Hadley and Marion Stuart Foundation, Mars Veterinary, generous private donors and the Flint Animal Cancer Center at Colorado State University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. MAF designed the annual GRLS owner survey and collected all survey data from the owners.
Funding was received from Morris Animal Foundation (D24 CA- 607). Data was received under Morris Animal Foundation Grant/Data Transfer Agreement (D23 CLP- 204).

Data availability

The datasets that support the findings of this study are freely available for researchers who request access to the Morris Animal Foundation's Data Commons at https://datacommons.morrisanimalfoundation.org/. All R analysis code is available in the MicroBatVet/MAF_GRLS Github Repository, [https://github.com/MicroBatVet/MAF_GRLS/tree/main].

Declarations

Ethics approval and consent to participate

As previously described by Guy et al. 2015 [15], all dog owners signed a consent form to participate in the GRLS. The GRLS was reviewed by an independent Animal Welfare Advisory Board of Morris Animal Foundation and approved the study protocol [15]. Ross University School of Veterinary Medicine (RUSVM) IACUC granted an exemption for the use of the anonymous GRLS dataset obtained from MAF GRLS.

Consent for publication

Not applicable.

Competing interests

D.R. is a 2023 Morris Animal Foundation Veterinary Student Scholar and was awarded a stipend to support his work on this project. S.H. is a member of Morris Animal Foundation's Animal Health Advisory Council.

Author details

[1]Department of Biomedical Sciences, Ross University School of Veterinary Medicine, Basseterre, Saint Kitts and Nevis. [2]College of Veterinary Medicine, Arkansas State University, Jonesboro, AR, USA.

Received: 5 February 2025 Accepted: 31 March 2025
Published online: 07 May 2025

References

1. Bronson RT. Variation in age at death of dogs of different sexes and breeds. Am J Vet Res. 1982;43:2057–9.
2. Adams VJ, Evans KM, Sampson J, Wood JLN. Methods and mortality results of a health survey of purebred dogs in the UK. J Small Anim Pract. 2010;51:512–24.
3. Dobson JM. Breed-predispositions to cancer in pedigree dogs. Int Sch Res Not. 2013;2013: 941275.
4. Inoue M, Hasegawa A, Hosoi Y, Sugiura K. A current life table and causes of death for insured dogs in Japan. Prev Vet Med. 2015;120:210–8.
5. Fleming JM, Creevy KE, Promislow DEL. Mortality in North American dogs from 1984 to 2004: an investigation into age-, size-, and breed-related causes of death. J Vet Intern Med. 2011;25:187–98.
6. Wise CF, Breen M, Stapleton HM. Canine on the Couch: The New Canary in the Coal Mine for Environmental Health Research. Environ Health. 2024;2:517–29.
7. Gardner HL, Fenger JM, London CA. Dogs as a model for cancer. Annu Rev Anim Biosci. 2016;4:199–222.
8. Burrai GP, Gabrieli A, Moccia V, Zappulli V, Porcellato I, Brachelente C, et al. A Statistical Analysis of Risk Factors and Biological Behavior in Canine Mammary Tumors: A Multicenter Study. Animals. 2020;10:1687.
9. Vazquez E, Lipovka Y, Cervantes-Arias A, Garibay-Escobar A, Haby MM, Queiroga FL, Velazquez C. Canine Mammary Cancer: State of the Art and Future Perspectives. Animals : an open access journal from MDPI. 2023;13(19):3147. https://doi.org/10.3390/ani13193147.
10. Schneider R, Dorn CR, Taylor D. Factors influencing canine mammary cancer development and postsurgical survival. J Natl Cancer Inst. 1969;43:1249–61.
11. Santos M, Marcos R, Faustino A. Histological study of canine mammary gland during the oestrous cycle. Reprod Domest Anim. 2010;45:e146–54.
12. Kent MS, Burton JH, Dank G, Bannasch DL, Rebhun RB. Association of cancer-related mortality, age and gonadectomy in golden retriever dogs at a veterinary academic center (1989–2016). PLoS ONE. 2018;13: e0192578.
13. Craig L. Cause of death in dogs according to breed: a necropsy survey of five breeds. J Am Anim Hosp Assoc. 2001;37:438–43.
14. History and future directions of the golden retriever lifetime study. https://www.morrisanimalfoundation.org/article/history-and-future-direction-golden-retriever-lifetime-study. Accessed 23 Nov 2024.
15. Guy MK, Page RL, Jensen WA, Olson PN, Haworth JD, Searfoss EE, et al. The Golden Retriever Lifetime Study: establishing an observational cohort study with translational relevance for human health. Philos Trans R Soc B Biol Sci. 2015;370:20140230.
16. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. J Thorac Dis. 2015;7(11):E537–40. https://doi.org/10.3978/j.issn.2072-1439.2015.10.63.
17. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1:2.
18. Moglia V, Johnson O, Cook G, de Kamps M, Smith L. Artificial intelligence methods applied to longitudinal data from electronic health records for prediction of cancer: a scoping review. BMC Med Res Methodol. 2025;25:1–17.
19. Wu X, Li W, Tu H. Big data and artificial intelligence in cancer research. Trends Cancer. 2024;10:147–60.
20. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. cancer inform. 2006. https://doi.org/10.1177/117693510600200030.
21. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform. 2023;24(2):bbad002. https://doi.org/10.1093/bib/bbad002.
22. Speiser JL, Wolf BJ, Chung D, Karvellas CJ, Koch DG, Durkalski VL. BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. Chemom Intell Lab Syst. 2019;185:122–34.
23. Lee HS, Kim J-H. The dog as an exercise science animal model: a review of physiological and hematological effects of exercise conditions. Phys Act Nutr. 2020;24:1.
24. Courcier EA, Thomson RM, Mellor DJ, Yam PS. An epidemiological study of environmental factors associated with canine obesity. J Small Anim Pract. 2010;51:362–7.
25. R Core Team. R: a language and environment for statistical computing. Viena, Austria: R Foundation for statistical computing; 2024.
26. RStudio Team. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. 2024. URL http://www.rstudio.com/.
27. Wickham, Hadley, Frnacois, Romain, Henry, Lionel, Muller, Kirill, Vaughan, Davis. dplyr: A Grammar of Data Manipulation. dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org.
28. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B, Krylov I. data.table: Extension of `data.frame`. R package version 1.17.0. 2025. https://CRAN.R-project.org/package=data.table.
29. Gorman B. mltools: Machine Learning Tools. R package version 0.3.5. 2018. https://CRAN.Rproject.org/package=mltools.
30. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.
31. RSBID: Resampling Strategies for Binary Imbalanced Datasets. https://github.com/dongyuanwu/RSBID?tab=readme-ov-file. Accessed 6 June 2024.
32. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008;28(5):1–26. https://doi.org/10.18637/jss.v028.i05.
33. Speiser JL. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J Biomed Inform. 2021;117: 103763.
34. Hooper SE, Hecker KG, Artemiou E. Using Machine Learning in Veterinary Medical Education: An Introduction for Veterinary Medicine Educators. Vet Sci. 2023;10(9):537. https://doi.org/10.3390/vetsci10090537.
35. Sarver AL, Makielski KM, DePauw TA, Schulte AJ, Modiano JF. Increased risk of cancer in dogs and humans: A consequence of recent extension of lifespan beyond evolutionarily determined limitations? Aging Cancer. 2022;3:3–19.
36. Pet Owner Resources. Veterinary Cancer Society. https://vetcancersociety.org/resources/pet-owners/pet-owner-resources/. Accessed 30 Nov 2024.
37. Labadie J, Swafford B, DePena M, Tietje K, Page R, Patterson-Kane J. Cohort profile: The Golden Retriever Lifetime Study (GRLS). PLoS ONE. 2022;17: e0269425.
38. Na H, Oliynyk S. Effects of physical activity on cancer prevention. Ann N Y Acad Sci. 2011;1229:176–83.
39. Rock CL, Thomson C, Gansler T, Gapstur SM, McCullough ML, Patel AV, et al. American Cancer Society guideline for diet and physical activity for cancer prevention. CA Cancer J Clin. 2020;70:245–71.
40. Physical Activity Guidelines for Americans, 2nd edition - Healthy People 2030 | odphp.health.gov. https://odphp.health.gov/healthypeople/tools-action/browse-evidence-based-resources/physical-activity-guidelines-americans-2nd-edition. Accessed 30 Nov 2024.
41. Speiser JL, Callahan KE, Ip EH, Miller ME, Tooze JA, Kritchevsky SB, et al. Predicting future mobility limitation in older adults: a machine learning analysis of health ABC Study Data. J Gerontol Ser A. 2022;77:1072–8.
42. Muti HS, Heij LR, Keller G, Kohlruss M, Langer R, Dislich B, et al. Development and validation of deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric cancer: a retrospective multicentre cohort study. Lancet Digit Health. 2021;3:e654–64.
43. Physical Activity - Healthy People 2030 | odphp.health.gov. https://odphp.health.gov/healthypeople/objectives-and-data/browse-objectives/physical-activity#cit1. Accessed 29 Nov 2024.
44. Baughman C, Norman K, Mukamal K. Adherence to American cancer society nutrition and physical activity guidelines among cancer survivors. JAMA Oncol. 2024;10:789–92.
45. Lee W. Exercise and Cancer Patients. Animal Cancer Foundation. 2023. https://acfoundation.org/exercise-and-cancer-patients/. Accessed 30 Nov 2024.
46. Lynch BM, Neilson HK, Friedenreich CM. Physical Activity and Breast Cancer Prevention. In: Courneya KS, Friedenreich CM, editors. Physical Activity and Cancer. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 13–42.
47. Thompson HJ, Westerlind KC, Snedden J, Singh M. Exercise intensity dependent inhibition of 1-methyl-l-nitrosourea induced mammary carcinogenesis in female F-344 rats. Carcinogenesis. 1995;16:1783–6.
48. Wang Q, Zhou W. Roles and molecular mechanisms of physical exercise in cancer prevention and treatment. J Sport Health Sci. 2021;10:201–10.
49. Westerlind KC, McCarty HL, Gibson KJ, Strange R. Effect of exercise on the rat mammary gland: implications for carcinogenesis. Acta Physiol Scand. 2002;175:147–56.
50. del Sáez MC, Barriga C, García JJ, Rodríguez AB, Ortega E. Exercise-induced stress enhances mammary tumor growth in rats: Beneficial effect of the hormone melatonin. Mol Cell Biochem. 2007;294:19–24.

51. Feb 16 JBU, Feb 16 2024 | 5 Minutes Updated:, Minutes 2024 | 5. Hydrotherapy for Dogs: A Growing Trend in Canine Physical Therapy. American Kennel Club. https://www.akc.org/expert-advice/health/hydrotherapy-for-dogs/. Accessed 30 Nov 2024.
52. US EPA O. 2017 National Water Quality Inventory Report to Congress. 2017. https://www.epa.gov/waterdata/2017-national-water-quality-inventory-report-congress. Accessed 30 Nov 2024.
53. Chau KNM, Carroll K, Li X-F. Swimming benefits outweigh risks of exposure to disinfection byproducts in pools. J Environ Sci. 2025;152:527–34.
54. Gouveia P, Felgueiras F, Mourão Z, Fernandes EDO, Moreira A, Gabriel MF. Predicting health risk from exposure to trihalomethanes in an Olympic-size indoor swimming pool among elite swimmers and coaches. J Toxicol Environ Health A. 2019;82:577–90.
55. Florentin A, Hautemanière A, Hartemann P. Health effects of disinfection by-products in chlorinated swimming pools. Second Eur PhD Stud Workshop Water Health Cannes. 2010;2011(214):461–9.
56. Braman SL, Peterson H, Elbe A, Mani E, Danielson C, Dahman C, et al. Urinary and household chemical exposures in pet dogs with urothelial cell carcinoma. Vet Comp Oncol. 2024;22:217–29.
57. Rafalko JM, Kruglyak KM, McCleary-Wheeler AL, Goyal V, Phelps-Dunn A, Wong LK, et al. Age at cancer diagnosis by breed, weight, sex, and cancer type in a cohort of more than 3,000 dogs: Determining the optimal age to initiate cancer screening in canine patients. PLoS ONE. 2023;18: e0280795.
58. German AJ, Blackwell E, Evans M, Westgarth C. Overweight dogs exercise less frequently and for shorter periods: results of a large online survey of dog owners from the UK. J Nutr Sci. 2017;6: e11.
59. Mao J, Xia Z, Chen J, Yu J. Prevalence and risk factors for canine obesity surveyed in veterinary practices in Beijing. China Prev Vet Med. 2013;112:438–42.
60. Marchi PH, Vendramini THA, Perini MP, Zafalon RVA, Amaral AR, Ochamotto VA, Da Silveira JC, Dagli MLZ, Brunetto MA. Obesity, inflammation, and cancer in dogs: Review and perspectives. Front Vet Sci. 2022;9:1004122. https://doi.org/10.3389/fvets.2022.1004122.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.