Variables associated with owner perceptions of the health of their dog: further analysis of data from a large international survey

Richard Barrett-Jolley, Alexander J. German*

Institute of Life Course and Medical Sciences, University of Liverpool, UK

Running Title: Variables associated with dog health

* Corresponding author

E-mail: ajgerman@liverpool.ac.uk

Abstract

In a recent study (doi: 10.1371/journal.pone.0265662), associations were identified between owner-reported dog health status and their diet, whereby those fed either raw food or a vegan diet were perceived to be healthier. However, the study was limited because possible confounding from variables not included in the analysis was not considered. The aim of the current study was to extend the earlier findings, using different modelling techniques and including multiple variables, to identify the most important predictors of owner-reported dog health.

From the original dataset, two binary health outcome variables were created: the 'any healthcare problem' distinguished those that were healthy ("no") from with those with illness of any severity, whilst the 'significant illness' variable distinguished those that were either healthy or had mild illness ("no") from those with significant or serious illness ("yes"). Associations between these health outcomes and both owner-animal metadata and healthcare variables were assessed using logistic regression and machine learning predictive modelling using XGBoost.

For the 'any health problem' outcome, best-fit models for both logistic regression (pseudo R^2 0.428, $P<0.001$) and XGBoost (AUC 0.838, 99%-CI: 0.794-0.882) contained the variables dog age, number of veterinary visits and receiving medication. Several other variables were also included in the XGBoost model, albeit having limited although their importance was limited. For the 'significant illness' outcome, dog age, veterinary visits and receiving medication were again the most important predictors for both logistic regression (R^2 0.349, $P<0.001$) and XGBoost (AUC 0.871, 99%-CI: 0.799-0.944), whilst location (UK) also featured in the latter. No other variable was of sufficient importance to be included in the final model.

The results of the current study conflict with previously-published research using the same dataset and suggest that diet the fed has limited impact on owner-perceived dog health status; instead, dog age, frequency of veterinary visits and receiving medication are most important.

Introduction

There is an increasing interest from owners in feeding unconventional diets, such as those utilising uncooked ingredients (so-called 'raw diets') and those utilising plant-based ingredients (either vegetarian or vegan diets). In a recent UK survey, 7% and approaching 1% of pet owners reportedly fed their dog either a raw or vegan diet, respectively [1]. Veganism is an increasingly popular food lifestyle choice in humans, with recent surveys indicating 2-3% [2], 4% [3] and 3% [4] of people in the UK, Europe and USA, respectively, to be vegan. Although still uncommon (<1% of owners), use of plant-based (vegan or vegetarian) diets for dogs is increasing in popularity [1], not least amongst owners who are vegan themselves [5]. Interestingly, the popularity of feeding raw diets is also increasing, with the same survey suggesting that 7% of the UK dog-owning population use this method [1], a proportion broadly similar to a recent international study, where 9% of dog owners exclusively fed raw food [6]. The main reason that owners cite for considering a plant-based diet is concern about animal welfare [5], whilst owners who feed raw diets believe these to be more 'natural' and healthier than commercial food despite an absence of evidence to support this [7]. Despite their increasing popularity, concerns remain regarding the safety of feeding unconventional foods; for both vegan [5,8] and raw [9-11] food, in terms of nutritional adequacy. For raw food, there are additional concerns about contamination either with pathogenic bacteria and bacteria resistant to multiple antimicrobials [11-13].

In recent studies, the safety of feeding unconventional diets for dogs has been examined, albeit using owner reports of health from questionnaires or surveys. In one study of 16,475 households that fed raw food, only 39 (0.2%) recalled a member of the household acquiring a pathogenic infection during the time they were using the food [14], perhaps, suggesting that risks to owners from bacterial contamination of raw food are uncommon. The health of dogs being fed unconventional diets has also been assessed. In one survey of 1,413 owners

conducted across North America, those feeding plant-based diets reported fewer health conditions in their dogs and a longer lifespan, compared with owners of dogs fed a meat-based diet [6]. In a second study [15], owner-reported health of dogs being fed various diet types was assessed. Compared with dogs fed a conventional diet, owners that fed either raw food or a vegan diet reported better health. However, besides being reliant on owner-reported health, rather than more objective health measures, no account was taken of possible confounding from other variables. To the authors' credit, the original study data were made available for use by other researchers (https://rebrand.ly/2020/pet-food-consumers-study), which not only included the diet and health data they analysed, but also other metadata including owner (e.g., country, career, education, age and gender) and animal (e.g., age, sex, neuter status and breed) variables, and healthcare-related variables (e.g., veterinary visits, use of medication, use of therapeutic diet). Therefore, the aim of the current study was to extend the earlier findings using different statistical techniques to create models that best predicted owner opinions of health, and to identify the relative importance of the variables that contributed to the final model. This included the use of ordinal and binary logistic regression and well as machine learning predictive modelling using a scalable tree boosting system (XGBoost; [16]).

Materials and Methods

Study design

A secondary analysis was conducted using data from a recent international survey of owner opinions on the health of their dog [15]. The open-access dataset from this study is accessible at https://rebrand.ly/2020/pet-food-consumers-study. The reason for conducting this further analysis was that many variables gathered were not analysed statistically, and that

the effect of possible confounding was not adequately tested, both of which were limitations acknowledged by the study authors [15].

Dataset used

Full details of survey design, methodology, information collected and category definitions have been reported previously [15]. Briefly, the original survey was designed using an online survey platform (https://www.onlinesurveys.ac.uk), with owners being asked to record details about their dog (including demographic data, main diet type and health characteristics) and themselves. Owner information used for the current study included location (country or continental region), setting (urban, rural or a mix of urban and rural), education level, qualifications achieved, household income, whether their occupation was in an animal-related career (including veterinarian, veterinary technician or nurse, animal breeder, animal trainer or pet industry worker), age, gender and diet. Animal data used for the current study included: age, sex, neuter status, breed group (e.g., toy, small, medium, large and giant) and main diet type (including conventional, raw, vegetarian, vegan, *in vitro* meat, insect, fungal or algal); health-related information used for the current study included owner impression of the health of their dog (healthy; generally healthy with minor or infrequent problems; significant or frequent problems; seriously ill), number of veterinary visits in the last 12 months, whether the dog had received medication (besides those used in preventive care e.g. routine vaccinations and endo- or ecto-parasite treatments) in the last 12 months, and whether the dog was on a therapeutic diet.

Statistical methods

Statistical software

All data pre-processing and statistical analyses were conducted using an online open-access statistical language and environment (R, version 4+) [17], with several packages including: 'aod' version 1.3.2 [18], 'broom' version 0.8.0 [19], 'car' version 3.0.13 [20], 'caret' version 6.0.93 [21], 'corrplot' version 0.92 [22], 'data,table' version 1.14.2 [23], 'dbscan' version 1.1-11 [24], 'DescTools' version 0.99.45 [25], 'devtools' version 2.4.3 [26], 'dplyr' version 1.0.9 [27], 'fmsb' version 0.7.3 [28], 'foreign' version 8.82 [29], 'generalhoslem' version 1.3.4 [30], 'ggplot2' version 3.3.6 [31], 'ggthemes' version 4.2.4 [32], 'glmtoolbox' version 0.1.3 [33], 'gofcat' version 0.1.2 [34], 'heatmaply' version 1.4.0 [35], 'Hmisc' version 4.7-0 [36], 'IHW' version 1.26.0 [37], 'largesampleehl' version 0.2.0 [38], 'MASS' version 7.3.57 [39]. 'mltools' version 0.3.5 [40], 'pROC' version 1.18.0 [41], 'readxl' version 1.4.0 [42], 'rsample' version 1.1.0 [43], 'reshape2' version 0.8.9 [44], 'ROCR' version 1.0-11 [45], 'rstatix' version 0.7.0 [46], 'showtext' version 0.9-5 [47]; 'smotefamily' version 1.3.1 [48], 'stringr' version 1.4.0 [49], 'tidyverse' version 1.3.1 [50], 'umap' version 0.8.2.0 [51], 'VGAM' version 1.1-6 [52] and xgboost version 1.6.0.1 [16]. Statistical reports of all data pre-processing and analysis are included in the supplementary information (S1-S6 Files), whilst all code is available online (https://github.com/RichardBJ/CanineHealth22).

Initial dataset pre-processing

Initially, data were imported into the statistical environment using the 'readxl' package [42], and then manipulated using the 'dplyr' package [27]. Columns were renamed for clarity and to improve compatibility with use in R (see below). Counts of all categorical data were examined, and categories with small groups were either removed or combined with others to exclude missing data or to reduce dataset imbalance. The variables included in the dataset were owner variables, animal variables (collectively owner-animal metadata), healthcare variables and health outcomes variables.

For owner variables, pre-processing included combining categories with small numbers in the *location* variable (e.g., 'Africa' [n=7], 'Asia' [n=25], 'other' [n=11] and 'South America' [n=47]) into a new 'other' category, and also removing the 'other' category [n=15] from the *settings* variable. The *owner education* variable was simplified by combining the 'basic' [n=15] and 'high school' [n=478] categories and combining the 'PG' [n=515] and 'PhD' [n=90] categories. A binary *animal career* variable was created from the *career* variable by combining the 'veterinarian' (n=125), 'veterinary technician or nurse' (n=54), 'animal breeder' (n=17), 'animal trainer' (n=140) and 'pet industry worker' (n=205) categories into a 'yes' category, whilst the remaining option ('none of the above' [n=2,137]) formed the 'no' category. The *owner income* variable was simplified by removing the 'prefer not to answer' category (n=184) and ordered, whilst the *owner age* variable was simplified (by combining the '18-19y' [n=19] and '20-29y' [n=391] categories into a new '<30y' category and combining the '60-69y' [n=364] and '>70y' [n=85] categories into a new '>60y' category) and then ordered. Given small numbers, the 'prefer not to answer' (n=17) and 'other '(n=4) categories were removed from *owner gender* variable, and the 'other' category (n=18) was removed from the *owner diet* variable. Further, binary variables were created for owners who reported consuming a vegan diet (yes vs. no) and owners who reported consuming either a vegan or vegetarian diet (yes or no).

For animal variables, pre-processing included removing data from dogs where age was reported as 'unsure' (n=3) or aged <1 year (n=26). In addition to having a continuous *dog age* variable, an ordinal variable was also created with data grouped into quintiles (1-3y, 3-5y, 5-7y, 7-9y, 9-20y). For the *dog diet* variable, the 'unsure' (n=8), 'insect-based' (n=5), 'meat-based (lab-grown)' (n=6) and 'mixture' (n=13) categories were all removed since numbers were again small. Further, binary variables were created for dogs consuming a vegan (yes vs.

no), vegan or vegetarian (yes or no) or a raw (yes vs. no) diet, and for dogs in the 'giant breed' category (yes vs. no) of the *breed size category* variable.

For healthcare variables, the 'unsure' category (n=13) was removed from the *veterinary visits* variable, and two ordinal variables were created; the first comprised 5 categories ('0 visits', '1 visit', '2 visits', '3 visits', '4+ visits'), whilst the second comprised 4 categories ('0 visits', '1 visit', '2 visits', '3+ visits' [n.b., '3 visits' and '4+ visits' categories combined]).

For the health outcomes variables, pre-processing included removing data where the owner was unsure of their dog's health (n=5). From this, an ordered *illness severity* variable was created with three categories ('none', 'mild', 'significant'); the 'none' and 'mild' categories contained the 'healthy' and 'generally healthy with minor or infrequent problems' categories, respectively whereas the 'significant' categories contained both the 'significant or frequent problems' and 'seriously ill' categories (given small numbers in the latter category). Two binary health status variables were also created: the first was an *any health problem* variable, where dogs classified as 'healthy' were classified as 'no', and all remaining categories ('generally healthy with minor or infrequent problems', 'significant or frequent problems' and 'seriously ill') were classified as 'yes'; the second was a *significant illness* variable, whereby dogs classified either as 'healthy' or 'generally healthy with minor or infrequent problems' were classified as 'no', and those classified as having 'significant or frequent problems' or being 'seriously ill' were classified as 'yes'.

As a final step, data were removed from owners that reported that they either played no role (n=15) or a lesser role (n=96) in decision making for their dog given uncertainty with reliability of these data. Ultimately, therefore, the final dataset comprised data from 2,211 owner and dog observations (Table 1).

Owner-dog metadata visualised by Uniform Manifold Approximation and Projection with Density-Based Spatial Clustering

Uniform Manifold Approximation and Projection (UMAP) [53] was used for initial dimension reduction (using the 'umap' package [51]), whilst density-based spatial clustering (DBSCAN, using the 'dbscan' package [24]) was applied to identify groupings within the owner-animal metadata [53]. With this analysis, each record (individual row of the dataset with one owner-dog combination) contributes one point in the UMAP. The following variables were included, all of which were coerced to numeric before analysis: dog and owner; diet, age and gender, together with dog-size and neuter-status, then owner; location, education and income.

Visualising associations amongst owner-animal metadata and healthcare variables

Kendall's *tau* [54] was used to conduct exploratory correlation using the R package 'rstatix' [46], and a correlation matrix [55] was plotted with the 'corrplot' package [22]. Data were first coerced into numerical, ordinal data, except for the *location* variable because there was no ordinal relationship amongst categories; instead, this variable was one-hot encoded, whereby a new binary variable was added for each category. *P* values are expressed as their negative logarithm (-Log*P*), whilst false discovery rate (FDR) was calculated by the Benjamini-Hochberg procedure using the 'IHW' package [37]. Hierarchical clustering was added with R stats::hclust using the Ward method [56]. Strengths of association were reported according to Cohen [57], whereby correlation coefficients of 0.1-0.3, 0.3-0.5 and 0.5-1.0 represented small, medium and large correlation, respectively.

Ordinal logistic regression

224 Ordinal logistic regression models were created using either the 'polr' function of the
225 'MASS' package [39] or the 'vglm' function of the 'VGAM' package [52]. The outcome
226 variable was *illness severity* (none = 0, mild =1, serious = 2), whilst all owner, animal and
227 healthcare variables were tested as predictor variables. Initially, simple ordinal regression was
228 used to test associations between illness severity and single predictor variables. Variables that
229 had been coded in different ways (e.g., *dog age* as both a continuous and ordinal variable, *dog*
230 *diet* with all categories or as separate binary variables [e.g. vegan, raw etc]), breed as both the
231 *breed size category* or just *giant breed* variables) were tested separately to determine which
232 fitted the data best based on the Bayesian Information Criterion (BIC; see below) [58]. An
233 'all-variable' multiple regression model was also built which contained all variables (owner-
234 animal metadata and healthcare variables), except for those coded in different ways, where the
235 option providing the best fit on simple regression was used.

236 A 'best fit' multiple regression model was also created which initially included variables
237 that met the threshold of $P < 0.2$ on simple regression. This model was refined in a backwards
238 and forwards stepwise fashion, with the BIC being used to select the model within the same
239 family with the best generalisability (a measure of its goodness of fit compared with its
240 complexity) [58]. With this approach, the existing model was repeatedly refined with addition
241 or removal of variables until the model with the smallest BIC was found, according to
242 previously published rules [59]. Models with interaction terms were also tested when these
243 were clinically relevant (e.g., between dog sex and neuter status, between owner and dog diet
244 categories and between location and setting). Results are reported as odds ratios (OR) with the
245 associated 99% confidence intervals (99%-CI). For variables that did not meet the proportional
246 odds assumption, separate OR and 99%-CI are reported for the different levels of illness
247 severity ('minor' and 'significant').

For model validation, the proportionality assumption was examined graphically, with partial proportional odds models being used when models include variables that did not meet this assumption. Influential datapoints were identified and assessed using Cook's distance; given that the data were from a secondary source, such datapoints were not removed, as their validity could not be checked. Possible multicollinearity in all models was explored using the variance inflation factor (VIF) for predictors with 1 degree of freedom [Df]) or generalised variance inflation factor (GVIF) and $GVIF^{(1/(2\times Df)}$ for predictors with >1 Df (e.g., polynomial variables) [60]. A rule of thumb was applied whereby multicollinearity was not considered to be a concern when VIF or GVIF was <4 or $GVIF^{(1/(2\times Df)}$ was <2. If necessary, multicollinearity was resolved by removing the variable with the greatest VIF or creating a new dummy variable by combining variables with high collinearity. Goodness of fit was tested by visual inspection of observed and expected results and 3 tests for goodness of fit: the Lipsitz goodness of fit test for ordinal response models, the Pulkstenis-Robinson chi-squared test and the Hosmer and Lemeshow test for large datasets. The coefficient of determination (pseudo-R^2) of the overall model was calculated using a modification of MacFadden's R^2 proposed by Ugba and Gertheiss [61].

Binary logistic regression

Binary logistic regression models were created using the 'glm' function in R. Both binary illness variables ('any health problem' and 'significant illness') were used separately as outcome variables, with the same predictor variables used as for ordinal logistic regression. Simple and multiple logistic models were created and refined using the same approach to that described for ordinal logistic regression. Influential datapoints were identified and assessed using Cook's distance, but not removed unless there was a good reason to do so. Possible multicollinearity in all models was assessed using the generalised variance inflation factor

(GVIF) and GVIF($1/(2\times Df)$) and was deemed to be acceptable when all values were <4 and <2 for GVIF and GVIF($1/(2\times Df)$), respectively. Issues with multicollinearity were resolved in the same way as for ordinal regression. Goodness of fit was tested by a visual inspection of observed and expected results and the Hosmer and Lemeshow test for large datasets. The coefficient of determination (pseudo-R^2) of the overall model was calculated using the method reported by Nagelkerke [62]. Results are again reported as OR and 99%-CI.

Machine learning predictive modelling

A scalable tree boosting system (XGBoost; "eXtreme Gradient Boosted" tree) was implemented in the statistical environment [16]. Data were pre-processed as described above and, to facilitate XGBoost, ranked variables used for the Kendall's tau correlation (as described above) were coerced to a numeric datatype, and then subdivided 70:30 into training and test sets. A model was then built on the training data without any involvement of the test set. After pre-processing, but before training, training data were augmented, as per standard practice for machine learning [63]. We trialled two augmentation methods: SMOTE [48] and a similar method where underrepresented classes were duplicated, but with simple addition of signal noise. Final models used the latter given that it performed better than SMOTE. Test data were left untouched. Augmented training data were then input to the XGBoost cross-validation method to tune optimal tree size/iterations and this was fed to the XGBoost training method. The model was then built and saved to file. Receiver operating characteristic (ROC) curves were calculated and plotted with the saved model and the previously unseen test dataset using the packages 'pROC' [41] and 'ROCR' [45]. Prediction accuracy was estimated by calculating the ROC area under the curve (AUC), which could range from 0 to 1; with this metric, a model that performed no better than chance would have an AUC of 0.5, and models predicting better than by chance would have AUC >0.5, with an AUC of 1.0 suggesting perfect prediction.

Different measures of the relative importance (e.g., feature importance, gain, cover and frequency parameters) of the variables that contributed to each of the final models were calculated and depicted graphically.

Results

Summary of owner-animal metadata

Details of the final dataset used for statistical analysis are shown in Table 1 and were similar to the data used in the previous study [15]. Briefly, owners represented a broad age range, with a relatively even distribution in education level, from high school to postgraduate. However, the majority (2,055, 92.9%) of respondents were female, whilst most responses were from the United Kingdom (1,609, 72.8%), with other European countries the next most common region represented. Almost a quarter of the respondents (493, 22.3%) defined themselves as vegan with a further 221 (10.0%) reporting to be vegetarian. A range of ages, breed sizes and sexes of dog were included in the final dataset (Table 1), with the majority of male (294/1169, 74.9%) and female (192/1042, 81.6%) dogs being neutered. Approximately a third (732, 33.1%) of dogs were fed a raw meat diet, whilst 32 (1.5%) and 295 (13.3%) were fed a vegetarian or vegan diet, respectively.

We used a UMAP projection and density-based clustering to create a simple overall visualisation of relationships in these owner-animal metadata (Fig 1; S1 Fig). Several clusters were identified which, one of which had a predominance of owners identifying as vegan (324/381 [85%] points in the cluster, 56% of all vegan owners), suggesting similarity in their owner-animal metadata. In this same cluster, there was a predominance (97%) of dogs fed either a vegan (336/383 [88%] points in the cluster, 100% of all vegan dogs) or vegetarian (35/383 [9%] points in the cluster, 100% of all vegetarian dogs) diets. Besides this, some clustering was also evident for location and education, but not for any of the other variables (S1 and S2 Figs).

Table 1.
Details of the final dataset from 2,211 owner-dog pairs, which was used for statistical analyses.

Variable [1]	Categories	Number	Percentage
Owner variables			
Location	United Kingdom	1609	72.8%
	Other European country	318	14.4%
	North America	119	5.4%
	Australia / New Zealand / Oceania	97	0.4%
	Other region	68	0.3%
Setting	Urban	780	35.3%
	Equally urban and rural	703	31.8%
	Rural	728	32.9%
Owner age (years)	<30	340	15.4%
	30-39	482	21.8%
	40-49	452	20.4%
	50-59	515	23.3%
	≥60	422	19.1%
Owner gender	Female	2055	92.9%
	Male	156	7.1%
Education	Basic or high school	414	18.7%
	College	636	28.8%
	Graduate	645	29.2%
	Postgraduate	516	23.3%
Income	Low	352	15.9%
	Medium	1350	69.2%
	High	329	14.9%
Animal-related Career	Yes	425	19.2%
	No	1786	80.8%
Owner diet [2]	Omnivore	901	40.7%
	Omnivore (restricted)	479	21.7%
	Pescatarian	117	5.3%
	Vegetarian	493	22.3%
	Vegan	221	10.0%
Animal Variables			
Age (years)	1-3	488	23.6%
	3-5	458	22.1%
	5-7	379	18.3%
	7-9	308	14.9%

	9-20	435	21.0%
Breed size category			
	Medium	859	38.9%
	Toy	52	2.4%
	Small	439	19.8%
	Large	764	34.5%
	Giant	97	4.4%
Sex	Female	1042	47.1%
	Male	1169	52.9%
Neuter status	Sexually intact	486	22.0%
	Neutered	1725	78.0%
Dog diet [2]	Conventional	1152	52.1%
	Raw	732	33.1%
	Vegetarian	32	1.5%
	Vegan	295	13.3%
Healthcare variables			
Veterinary visits	None	399	18.0%
	1	873	39.5%
	2	473	21.4%
	3	185	8.4%
	4 or more	281	12.7%
On medication	No	1331	60.2%
	Yes	880	39.8%
On therapeutic food	No	2111	95.5%
	Yes	100	4.5%

[1] Definitions of the different categories are given in the original study [15]. [2] Please see the footnote to Table 1 for details of how owner and dog diets were assigned.

333

334 **Fig 1. Visualisation of owner metadata by Uniform Manifold Approximation and Projection with**
335 **Density-Based Spatial Clustering.** Owner-pet metadata were pre-processed with all factors as
336 numeric and subject to dimension reduction with the UMAP projection technique. Owner variables
337 included in this visualisation were *diet*, *sex*, *location*, *education* and *income*; animal variables included
338 were *age*, *sex*, *neuter status*, *diet* and *breed size category*. Healthcare variables were not included. Each
339 individual row of the data (owner-dog combination) contributes one UMAP x and y coordinate,
340 analogous to PC1 and PC2 in a principal component analysis. Four or five distinct clusters were evident.
341 (a) Points colour-coded by *owner diet*. Here, owners consuming a vegan diet predominated (324/381,
342 85%) in one cluster. (b) Points colour-coded by *dog diet*. Here, dogs fed either a vegan (336/336, 100%)
343 or vegetarian (35/35, 100%) diet predominated in one cluster, which was the same cluster from which
344 there was a predominance of vegan owners.

345

346

347 Summary of healthcare variables

348 Details of healthcare variables are given in Table 1. Most (1,812; 82.0%) dogs had
349 visited their veterinarian at least once in the last year, whilst 880 (39.8%) and 100 (4.5%) dogs
350 had received medication or had been switched to a therapeutic diet, respectively.

351

352 Associations amongst owner-animal metadata and healthcare 353 variables

354 Details of Kendall's tau correlations are shown in Fig 2, whilst datasets of all correlation
355 coefficients and -Log*P* are presented in the supplementary information (S1 and S2 Tables,
356 respectively). Several associations were identified including a large positive association
357 between owners and dogs consuming a *vegan diet* (Kendall's tau 0.67, -Log*P 201*), a small

358 positive association been UK location and dogs being fed a raw diet (Kendall's tau 0.25, -Log*P*
359 29), and a small negative association between *owner education* and the dogs consuming a *raw*
360 *diet* (Kendall's tau -0.11, -Log*P* 6.9). Positive associations were also evident amongst
361 healthcare variables (*receiving medication* and *veterinary visits*, Kendall's tau 0.52, -Log*P*
362 148.71; medication and use of a therapeutic diet, Kendall's tau 0.17, -Log*P* 14.16; having been
363 *switched to a therapeutic diet* and *veterinary visits*, Kendall's tau 0.16, -Log*P* 15.15).
364
365
366 **Fig 2. Naïve correlation amongst owner-animal metadata variables.** Data were first coerced into
367 numerical, ordinal data, except for the Location variable because there was no ordinal, where one-hot
368 encoding was used. Correlations used the Kendall's tau and for associations where the false discovery
369 rate was <0.05, the strength of correlation is indicated by the size of the inset circle, with positive and
370 negative associations coloured blue and red, respectively. Not surprisingly, a large blue circle is evident
371 when variables are correlated with themselves. Hierarchical clustering, using the Ward method [56], is
372 shown on the left side of the figure. The log *P*-values and correlation coefficients for all associations
373 are included in full in the supplementary tables (S1 Table, S2 Table).
374
375

376 Variables associated with owner-reported *illness severity*

377 Of the 2,211 dogs in the final dataset, owner-reported illness severity was 'none' in 1,375
378 dogs (62.2%), 'mild' in 714 dogs (32.3%) and serious in 122 dogs (5.5%). This variable was
379 assessed using ordinal logistic regression. The proportional odds assumption was not met for
380 several variables (e.g., *veterinary visits*, whether the dog was *receiving medication* and whether
381 the dog was a *giant breed*), where partial proportional odds logistic regression was instead used.
382 Results of simple ordinal logistic regression are presented in Table 2, whilst the all-variable
383 multiple regression model is shown in S3 Fig. Pseudo-R^2 for this all-variable model was 0.599

384 (P<0.001), and BIC was 2840. The following variables met the threshold for initial inclusion
385 in best-fit multiple regression modelling: *location*, *owner sex*, *animal-related career*, *owner*
386 *diet*, *dog age*, *breed size category*, *giant breed*, *dog diet*, *veterinary visits* (as a 4-category
387 ordinal variable), whether the dog was *receiving medication* and whether the dog was *switched*
388 *to a therapeutic food* (Table 2). After refinement by backward and forwards stepwise
389 elimination, the model with the best generalisability contained the following variables: *dog age*,
390 *veterinary visits*, *receiving medication* and having been *switched to a therapeutic food* (Fig 3,
391 S3 Table). Pseudo-R^2 for this best-fit model was 0.575 (P<0.001) and BIC was 2672.
392 Inclusion of any of the other variables identified on simple regression, either individually or in
393 combinations and with interactions, led to models with worse generalisability. Goodness-of-
394 fit was deemed to be acceptable based on a comparison of observed versus expected results
395 and by the Pulkstenis-Robinson Chi-squared test (P=1.000), but goodness-of-fit was worse
396 when tested with either the Lipsitz test (P=0.052) or Hosmer-Lemeshow test (P=0.038).
397 Therefore, two binary variables were created, an *any health problem* variable and a *significant*
398 *illness* variable, which could then be separately be assessed by binary logistic regression and
399 machine learning predictive modelling, as described below.
400

401 **Table 2.**
402 Results of simple ordinal logistic regression analyses examining associations between owner, animal and veterinary variables and *illness severity*
403 as reported by the owner.
404

Variable [1]	Odds ratio	99% confidence interval	*P*-value	Pseudo-R^2	BIC
Owner variables					
Location				0.0044	3667
United Kingdom	Ref	---	---		
Other European country	1.142	0.827-1.569	0.284		
North America	1.402	0.858-2.262	0.071		
Australia / New Zealand / Oceania	1.434	0.832-2.435	0.082		
Other region	0.959	0.482-1.825	0.870		
Setting				0.0018	3655
Urban	Ref	---	---		
Equally urban and rural	0.880	0.672-1.152	0.222		
Rural	0.852	0.649-.118	0.130		
Owner age (years)				0.0043	3667
<30	Ref	---	---		
30-39	1.054	0.733-1.152	0.707		
40-49	0.907	0.626-1.316	0.499		
50-59	0.784	0.544-1.132	0.871		
≥60	0.853	0.583-1.247	0.279		
Owner sex				0.0029	3646
Female	Ref	---	---		
Male	0.697	0.433-1.094	0.044		
Education				0.0012	3664
Basic or high school	Ref	---	---		
College	0.843	0.607-1.171	0.180		
Graduate	0.894	0.645-1.241	0.377		
Postgraduate	0.889	0.631-1.254	0.378		
Income				0.0019	3655

	Low	Ref	---	---		
	Medium	0.846	0.625-1.150	0.157		
	High	0.784	0.525-1.169	0.177		
Animal-related Career					0.0030	3646
	No	Ref	---	---		
	Yes	1.257	0.949-1.658	0.035		
Owner diet [2]					0.0080	3662
	Omnivore	Ref	---	---		
	Omnivore (restricted)	1.247	0.926-1.675	0.055		
	Pescatarian	1.754	1.072-2.485	0.003		
	Vegetarian	1.048	0.861-1.885	0.682		
	Vegan	1.278	0.778-1.410	0.106		
Animal Variables						
Age (years)					0.2186	3325
	1-3	Ref	---	---		
	3-5	1.122	0.776-1.626	0.419		
	5-7	1.345	0.918-1.970	0.045		
	7-9	2.283	1.550-3.370	<0.001		
	9-20	4.074	2.866-5.829	<0.001		
Breed size category					0.0050	3666
	Medium	Ref	---	---		
	Toy	1.120	0.561-2.453	0.521		
	Small	1.317	0.970-1.784	0.020		
	Large	1.202	0.924-1.564	0.071		
	Giant	1.420	0.803-2.467	0.106		
Giant breed [3]					0.0025	3654
	No	Ref	---	---		
	Yes (minor)	1.162	0.706-1.912	0.438		
	Yes (significant)	2.055	0.903-4.675	0.024		
Sex					0.0000	3650
	Female	Ref	---	---		

Male	1.001	0.801-1.252	0.988		
Neuter status				0.0214	3618
Sexually intact	Ref	---	---		
Neutered	1.853	1.393-2.486	<0.001		
Dog diet [2]				0.0261	3627
Conventional	Ref	---	---		
Raw	0.586	0.454-0.753	<0.001		
Vegetarian	1.048	0.403-2.589	0.895		
Vegan	0.567	0.394-0.805	<0.001		
Dog on vegan diet [2]				0.0054	3642
No	Ref	---	---		
Yes	0.689	0.485-0.967	0.005		
Dog on raw diet [2]				0.0142	3629
No	Ref	---	---		
Yes	0.651	0.509-0.829	<0.001		
Healthcare variables					
Veterinary visits [3]				0.3032	3191
None	Ref	---	---		
1 (minor)	1.657	1.096-2.507	0.002		
2 (minor)	6.107	3.974-9.386	<0.001		
3 or more (minor)	12.903	8.286-20.092	<0.001		
1 (significant)	0.913	0.221-3.775	0.869		
2 (significant)	2.938	0.783-11.022	0.036		
3 or more (significant)	18.862	5.693-62.493	<0.001		
Prescribed medication by vet [3]				0.378	3020
No	Ref	---	---		
Yes (minor)	10.258	7.934-13.263	<0.001		
Yes (significant)	21.639	8.942-52.367	<0.001		
Switched to therapeutic food				0.0560	3566
No	Ref	---	---		
Yes	6.217	3.726-10.443	<0.001		

[1] Definitions of the different categories are given in the original study [15]. [2] Please see the footnote to Table 1 for details of how owner and dog diets were assigned. [3] For variables that did not meet the proportional odds assumption, separate odds ratios and 99% confidence intervals are reported for the different levels of *illness severity* ('minor' and 'severe'). BIC: Bayesian information criterion, with models having the best fit having lower BIC [58,59]; BIC values can only be compared within the same family of models.

411 **Fig 3. Final ordinal logistic regression model of *illness severity*. Results of the best-fit ordinal**
412 **logistic regression model of illness severity.** The dots represent the odds ratio for each variable, whilst
413 the bars represent 99% confidence intervals (99%-CI). Variables where the 99%-CI range does not
414 include1.0 (vertical dotted line) are depicted in red, whilst those that include 1.0 are depicted in blue.
415 Note the logarithmic scale for the X-axis. *Receiving medication* and *veterinary visits* did not meet the
416 partial proportional odds assumption and, therefore, separate odds ratios and 99% confidence intervals
417 are reported for the different levels of illness severity ('minor' and 'severe').
418
419

420 Owner, animal and healthcare variables associated with the *any*
421 *health problem* binary variable

422 Results of simple binary logistic regression are presented in Table 3, whilst the all-
423 variable multiple regression model is shown in S4 Fig. Pseudo-R^2 for this all-variable model
424 was 0.453 (P<0.001), and BIC was 2231. The following variables met the threshold for initial
425 inclusion in best-fit multiple regression modelling: *location*, *owner diet*, *dog age*, *breed size*
426 *category*, *giant breed*, *dog diet*, *veterinary visits* (as a 5-category ordinal variable), whether the
427 dog was *receiving medication* and being *switched to a therapeutic food*. After refinement by
428 backwards and forwards stepwise elimination, the model with the best generalisability
429 contained the following variables: *dog age*, *veterinary visits*, *receiving medication* and having
430 been *switched to a therapeutic food* (Fig 4, S4 Table). Pseudo R^2 for this best-fit model was
431 0.428 (P<0.001) and BIC was 2058. Inclusion of any of the other variables identified on simple
432 regression, either individually or in combinations and with interactions, led to models with
433 worse generalisability. Goodness of fit was deemed to be reasonable based on a visual
434 comparison of observed versus expected results and the Hosmer-Lemeshow test (P=0.066).

435 **Table 3.**
436 Results of simple binary logistic regression analyses examining associations between owner, animal and veterinary variables and the *any health*
437 *problem* binary (the presence of any illness as reported by the owner).

Variable [1]	Odds ratio	99% confidence interval	*P*-value	Pseudo-R^2	BIC
Owner variables					
Location				0.0032	2966
United Kingdom	Ref	---	---		
Other European country	1.116	0.804-1.540	0.385		
North America	1.383	0.840-2.262	0.090		
Australia / New Zealand / Oceania	1.372	0.791-2.356	0.134		
Other region	0.940	0.470-1.804	0.810		
Setting				0.0015	2953
Urban	Ref	---	---		
Equally urban and rural	0.874	0.665-1.148	0.204		
Rural	0.859	0.651-1.132	0.156		
Owner age (years)				0.0043	2964
<30	Ref	---	---		
30-39	1.021	0.704-1.481	0.888		
40-49	0.899	0.616-1.314	0.469		
50-59	0.753	0.519-1.093	0.050		
≥60	0.833	0.566-1.225	0.221		
Owner sex				0.0027	2943
Female	Ref	---	---		
Male	0.692	0.429-1.089	0.041		
Education				0.0009	2962
Basic or high school	Ref	---	---		
College	0.861	0.616-1.203	0.248		
Graduate	0.888	0.637-1.239	0.358		
Postgraduate	0.884	0.624-1.253	0.361		
Income				0.0016	2953

	Low	Ref	---	---		
	Medium	0.842	0.619-1.151	0.154		
	High	0.789	0.525-1.184	0.133		
Animal-related Career					0.0023	2958
	No	Ref	---	---		
	Yes	1.235	0.929-1.638	0.054		
Owner diet [2]					0.0073	2959
	Omnivore	Ref	---	---		
	Omnivore (restricted)	1.210	0.896-1.632	0.102		
	Pescatarian	1.856	1.114-3.095	0.002		
	Vegetarian	1.030	0.825-1.824	0.178		
	Vegan	1.230	0.762-1.391	0.794		
Animal Variables						
Age (years)					0.0873	2658
	1-3	Ref	---	---		
	3-5	1.116	0.769-1.619	0.449		
	5-7	1.329	0.905-1.953	0.056		
	7-9	2.283	1.542-3.393	<0.001		
	9-20	4.004	2.793-5.782	<0.001		
Breed size category					0.0044	2964
	Medium	Ref	---	---		
	Toy	1.275	0.588-2.679	0.404		
	Small	1.346	0.986-1.835	0.014		
	Large	1.178	0.903-1.537	0.114		
	Giant	1.321	0.745-2.308	0.202		
Giant breed					0.0003	2947
	No	Ref	---	---		
	Yes	1.162	0.667-1.991	0.477		
Sex					0.0002	2947
	Female	Ref	---	---		
	Male	1.015	0.810-1.274	0.861		

Neuter status					0.0190	2917
	Sexually intact	Ref	---	---		
	Neutered	1.849	1.387-2.486	<0.001		
Dog diet [1]					0.0222	2927
	Conventional	Ref	---	---		
	Raw	0.593	0.459-0.765	<0.001		
	Vegetarian	1.007	0.384-2.659	0.984		
	Vegan	0.568	0.394-0.811	<0.001		
Dog on vegan diet [2]					0.0049	2940
	No	Ref	---	---		
	Yes	0.689	0.483-0.969	0.006		
Dog on raw diet [2]					0.0118	2928
	No	Ref	---	---		
	Yes	0.660	0.515-0.843	<0.001		
Healthcare variables						
Veterinary visits					0.2451	2531
	None	Ref	---	---		
	1	1.657	1.106-2.535	0.002		
	2	6.107	4.012-9.494	<0.001		
	3	8.311	4.949-14.228	<0.001		
	4 or more	18.050	10.978-30.474	<0.001		
Prescribed medication by vet					0.3242	2346
	No	Ref	---	---		
	Yes	10.259	7.904-13.393	<0.001		
Switched to therapeutic food					0.0450	2874
	No	Ref	---	---		
	Yes	6.729	3.652-13.371	<0.001		

[1] Definitions of the different categories are given in the original study [15]. [2] Please see the footnote to Table 1 for details of how owner and dog diets were assigned. BIC: Bayesian information criterion, with models having the best fit having lower BIC [58,59]; BIC values can only be compared within the same family of models.

Fig 4. Final binary logistic regression model of variables associated with the *any health problem* binary. Results of the best-fit binary logistic regression model of the *any health problem* binary, which included the mild, significant and serious categories of the *illness severity* variable (Fig 3). The dots represent the odds ratio for each variable, whilst the bars represent 99% confidence intervals (99%-CI). Variables where the 99%-CI range does not include1.0 (vertical dotted line) are depicted in red, whilst those that include 1.0 are depicted in blue. Note the logarithmic scale for the X-axis.

The ability of owner, animal and healthcare variables to predict the *any health problem* variable was further assessed using machine learning predictive modelling using the XGBoost algorithm. Two separate models were created, one containing all variables, whilst the second was a reduced model containing just the owner-animal metadata (i.e., without healthcare variables). Based on ROC analysis, the prediction accuracy of the all-variable model was reasonable (AUC 0.838, 99%-CI: 0.794-0.882; Fig 5a). The XGBoost procedure also enables the most important variables contributing to the prediction to be identified in order of importance and, for the all-variable model (Fig 5b), the variables that most strongly predicted the *any health problem* binary were (in order): *veterinary visits*, *dog age* and *receiving medication*, followed by *owner age*, *owner education* and then *breed size category*; other variables followed these including *dog diet*, *dog sex*, *owner income* and *location*. The order of importance of predictor variables was broadly similar in the reduced model although, based on ROC analysis, prediction accuracy was poor (AUC 0.665, 99%-CI: 0.607-0.723, Fig 6a). In this reduced model, *dog age* was overwhelmingly the most important predictor, with *owner age*, *education level*, *dog size* and *dog sex* being next; *owner gender*, *location*, *owner income* and *dog diet* followed these (Fig 6b).

468 **Fig 5. All-variable XGBoost model on the *any health problem* binary.** (a) Receiver operating
469 characteristic curve of a prediction model containing all variables (owner, animal and healthcare). This
470 shows the increasing true positive and false positive rates, with decrease of the threshold probability for
471 prediction of any health issue. Prediction accuracy was good as assessed by ROC analysis (area under
472 curve 0.838, 99%-CI: 0.794-0.882). Acceptance threshold is indicated by the colour bar on the right-
473 hand side and shown at discrete points on the curve. For example, with a threshold probability of 0.200,
474 the model correctly predicts approximately 80% of dogs where the *any health problem* binary was
475 classified as 'yes' reported, but with an approximately 30% false-negative rate. (b) Graph depicting the
476 relative contribution of predictor variables to the all-variable XGBoost model for the 'any health
477 problem' binary. Variables are organised in order of importance, from left to right, whilst variables
478 that are not shown did not contribute to the final model. Gain, cover frequency and importance are
479 different measures of the contribution of each feature in the model.
480
481 **Fig 6. Reduced XGBoost model on the *any health problem* binary.** (a) Receiver operating
482 characteristic curve of a reduced prediction model only containing owner and animal variables. This
483 shows the increasing true positive and false positive rates, with decrease of the threshold probability for
484 prediction of any health issue. Prediction accuracy was moderate, as assessed by ROC analysis (AUC
485 0.665, 99%-CI: 0.607-0.723). Acceptance threshold is indicated by the colour bar on the right-hand
486 side and shown at discrete points on the curve. For example, with a threshold probability of 0.200, the
487 model correctly predicts approximately 73% of dogs where the *any health problem* binary was
488 classified as 'yes' reported, but with an approximately 50% false-negative rate. The fact that the
489 prediction thresholds are all low (< 0.5) shows that this model struggles to predict health issues. (b)
490 Graph depicting the relative contribution of predictor variables to the all-variable XGBoost model for
491 the 'any health problem' binary. Variables are organised in order of importance, from left to right,
492 whilst variables that are not shown did not contribute to the final model. Gain, cover frequency and
493 importance are different measures of the contribution of each feature in the model.
494

Owner, animal and healthcare variables associated with the *significant illness* binary variable

Results of simple binary logistic regression are presented in Table 4, whilst the all-variable multiple regression model is shown in S5 Fig. Pseudo-R^2 for this all-variable model was 0.411 ($P<0.001$), and BIC was 893. The following variables met the threshold for initial inclusion in best-fit multiple regression modelling: *location*, *owner sex*, *owner diet*, *dog age*, *breed size category*, *giant breed*, *neuter status*, *dog diet*, *veterinary visits* (as a 5-category ordinal variable), *receiving medication* and whether the dog was *switched to a therapeutic food*. After refinement by backwards and forwards stepwise elimination of variables, the model with the best generalisability contained the following variables: *dog age*, *veterinary visits* and *receiving medication* (Fig 7, S5 Table). Pseudo R^2 for this best fit model was 0.349 ($P<0.001$) and BIC was 708. Once again, inclusion of any of the other variables identified on simple regression, either individually or in combinations and with interactions, led to models with worse generalisability. Goodness of fit was deemed to be acceptable based on a comparison of observed versus expected results and the Hosmer-Lemeshow test ($P=0.113$).

510 **Table 4.**

511 Results of simple binary logistic regression analyses examining associations between owner, animal and healthcare variables and the *significant*
512 *illness* binary variable ('significant' or 'serious illness' as reported by the owner).

Variable [1]	Odds ratio	99% confidence interval	*P*-value	Pseudo-R^2	BIC
Owner variables					
Location				0.0073	977
United Kingdom	Ref	---	---		
Other European country	1.459	0.737-2.696	0.130		
North America	1.606	0.555-3.788	0.195		
Australia / New Zealand / Oceania	2.007	0.689-4.783	0.059		
Other region	1.227	0.229-3.940	0.699		
Setting				0.0019	966
Urban	Ref	---	---		
Equally urban and rural	0.768	0.416-1.393	0.258		
Rural	0.955	0.540-1.679	0.833		
Owner age (years)				0.0037	979
<30	Ref	---	---		
30-39	1.488	0.681-3.486	0.204		
40-49	0.967	0.413-2.436	0.968		
50-59	1.253	0.564-2.960	0.478		
≥60	1.114	0.471-2.734	0.749		
Owner sex				0.0005	959
Female	Ref	---	---		
Male	0.793	0.239-1.966	0.559		
Education				0.0049	971
Basic or high school	Ref	---	---		
College	0.636	0.303-1.334	0.112		
Graduate	0.960	0.494-1.919	0.877		
Postgraduate	0.954	0.470-1.959	0.863		
Income				0.0013	966

	Low	Ref	---	---		
	Medium	0.882	0.481-1.735	0.612		
	High	0.717	0.285-1.721	0.333		
Animal-related Career					0.0039	956
	No	Ref	---	---		
	Yes	1.466	0.821-2.505	0.076		
Owner diet [2]					0.0092	975
	Omnivore	Ref	---	---		
	Omnivore (restricted)	1.735	0.921-3.249	0.023		
	Pescatarian	1.228	0.319-3.488	0.649		
	Vegetarian	1.893	0.833-4.017	0.035		
	Vegan	1.316	0.665-2.543	0.288		
Animal Variables						
Age (years)					0.0554	902
	1-3	Ref	---	---		
	3-5	1.322	0.497-3.633	0.461		
	5-7	1.712	0.655-4.678	0.150		
	7-9	2.537	1.007-6.777	0.011		
	9-20	4.961	2.302-12.100	<0.001		
Breed size category					0.0119	973
	Medium	Ref	---	---		
	Toy	0.412	0.005-2.835	0.386		
	Small	1.056	0.502-2.122	0.843		
	Large	1.504	0.859-2.668	0.062		
	Giant	2.417	0.845-5.941	0.018		
Giant breed					0.0047	956
	No	Ref	---	---		
	Yes	2.055	0.753-4.667	0.038		959
Sex					0.0009	
	Female	Ref	---	---		
	Male	0.855	0.528-1.384	0.401		

Neuter status					0.0086	953
	Sexually intact	Ref	---	---		
	Neutered	1.923	0.999-4.150	0.017		
Dog diet [2]					0.0168	962
	Conventional	Ref	---	---		
	Raw	0.487	0.260-0.861	0.002		
	Vegetarian	1.368	0.180-5.281	0.612		
	Vegan	0.561	0.227-1.189	0.068		
Dog on vegan diet [2]					0.0019	958
	No	Ref	---	---		
	Yes	0.696	0.285-1.448	0.244		
Dog on raw diet [2]					0.0113	951
	No	Ref	---	---		
	Yes	0.531	0.287-0.925	0.005		
Healthcare variables						
Veterinary visits					0.2576	775
	None	Ref	---	---		
	1	0.913	0.233-4.499	0.869		
	2	2.938	0.872-13.619	0.036		
	3	6.451	1.810-30.717	<0.001		
	4 or more	29.214	10.248-125.903	<0.001		
Prescribed medication by vet					0.2027	798
	No	Ref	---	---		
	Yes	21.641	9.676-60.718	<0.001		
Switched to therapeutic food					0.04332	926
	No	Ref	---	---		
	Yes	5.672	2.789-10.879	<0.001		

513 [1] Definitions of the different categories are given in the original study [15]. [2] Please see the footnote to Table 1 for details of how owner and dog
514 diets were assigned. BIC: Bayesian information criterion, with models having the best fit having lower BIC [58,59]; BIC values can only be
515 compared within the same family of models.
516

517 **Fig 7. Final binary logistic regression model of variables associated with the *significant illness***
518 **binary.** Results of the best-fit binary logistic regression model of the *significant illness* binary variable,
519 which included the 'significant' and 'serious' categories of the illness severity variable (Fig 3). The
520 dots represent the odds ratio for each variable, whilst the bars represent 99% confidence intervals (99%-
521 CI). Variables where the 99%-CI range does not include1.0 (vertical dotted line) are depicted in red,
522 whilst those that include 1.0 are depicted in blue. Note the logarithmic scale for the X-axis.
523
524
525 Using machine learning predictive modelling with XGBoost, all-variable and reduced
526 (only owner and animal variables) models were created (S2 File). Based on ROC analysis, the
527 prediction accuracy of the all-variable model was reasonable (AUC 0.871, 99%-CI: 0.799-
528 0.944; Fig 8a). In order of importance (Fig 8b), the variables that most strongly predicted the
529 *significant illness* binary were: *veterinary visits*, *receiving medication*, followed by *dog age*
530 and *location (UK)*; none of the other variables were of sufficient importance to be included in
531 the final model. As for the *any health problem* binary, ROC analysis suggested that prediction
532 accuracy of the reduced model was poor (AUC 0.678, 99%-CI: 0.561-0.795, Fig 9a). Once
533 again, *dog age* was overwhelmingly the most important predictor; *owner age*, *education*,
534 *location*, *dog size* and *dog sex* were next, whilst other variables followed these albeit with very
535 low importance (Fig 9b).
536
537
538 **Fig 8. All-variable XGBoost model on the 'significant illness' binary.** (a) Receiver operating
539 characteristic curve of a prediction model containing all variables (owner, animal and healthcare). This
540 shows the increasing true positive and false positive rates, with decrease of the threshold probability for
541 prediction of any health issue. Prediction accuracy was good, as assessed by ROC analysis (area under
542 curve 0.871, 99%-CI: 0.799-0.944). Acceptance threshold is indicated by the colour bar on the right-

543 hand side and shown at discrete points on the curve. For example, with a threshold probability of 0.345,
544 the model correctly predicts approximately 82% of dogs where the 'significant' binary was classified
545 as 'yes' reported, but with an approximately 27% false-negative rate. (b) Graph depicting the relative
546 contribution of predictor variables to the all-variable XGBoost model for the *any health problem* binary.
547 Variables are organised in order of importance, from left to right, whilst variables that are not shown
548 did not contribute to the final model. Gain, cover frequency and importance are different measures of
549 the contribution of each feature in the model.
550
551 **Fig 9. Reduced XGBoost model on the 'significant illness' binary.** (a) Receiver operating
552 characteristic curve of a reduced prediction model only containing owner and animal variables. This
553 shows the increasing true positive and false positive rates, with decrease of the threshold probability for
554 prediction of any health issue. Prediction accuracy was moderate, as assessed by ROC analysis (area
555 under curve 0.678, 99%-CI: 0.561-0.795). Acceptance threshold is indicated by the colour bar on the
556 right-hand side, and also shown at discrete points on the curve. For example, with a threshold
557 probability of 0.148, the model correctly predicts approximately 63% of dogs where the *significant*
558 *illness* binary was classified as 'yes' reported, but with an approximately 33% false-negative rate. (b)
559 Graph depicting the relative contribution of predictor variables to the all-variable XGBoost model for
560 the 'any health problem' binary. Variables are organised in order of importance, from left to right,
561 whilst variables that are not shown did not contribute to the final model. Gain, cover frequency and
562 importance are different measures of the contribution of each feature in the model.
563

Discussion

In the current study, we examined associations between owner perceptions of the health of their dog, and a range of owner-related, animal-related and healthcare variables. We utilised data from a previous study, which had used a questionnaire to gather the opinions of owners about the health of their dogs [15]. The primary aim of that study was to examine associations between owner opinions of dog health status and the type of diet that owners predominantly fed, with the key finding being that health was positively-associated with feeding either a vegan or raw diet, compared with other diet types (including conventional and vegetarian). However, other information gathered in the questionnaire was not assessed and no account was taken of possible confounding amongst variables. Therefore, to extend the earlier findings, we used a combination of statistical methods to analyse multiple variables concurrently to create models that best predicted owner opinions of health and illness in their dog, and to determine the relative importance of variables contributing to the final models. In using both multiple logistic regression and machine learning predictive modelling, we hoped to maximise the benefits of both approaches, whilst minimising disadvantages. Multiple logistic regression is a well-known, widely available technique that is familiar to most people and has been extensively used in veterinary studies including those in dogs [64-67]. Procedures for model selection are well established, such as those based on BIC [58), which can reduce the risk of model overfitting and can take relevant prior knowledge into account. For example, in the current study, as part of our exploratory analyses, we tested interactions with a possible clinical (e.g., between *veterinary visits* and *receiving medication*), biological (between *dog sex* and *neuter status*) or epidemiological (between geographical *location* and *setting*) and psychological (between diet choice for owners and dogs). Of course, one disadvantage of the use of such prior knowledge is that it could introduce bias into the selection process. Other disadvantages include the fact that logistic regression makes certain statistical assumptions (e.g.,

589 independence of errors, the requirement for a linear relationship between the logit of the
590 outcome and predictors, absence of multicollinearity and lack of strong influential outliers;
591 [68]) which, if not met, could lead to poor predictive capability. Key advantages of machine
592 learning methods are the fact that they make fewer prior assumptions about data distribution
593 and, therefore, can better model complex, nonlinear relationships between the outcome and any
594 predictors even when there is significant noise [69]. However, there are disadvantages
595 including requiring a large dataset, the time and resources needed for computation, challenges
596 with interpretation of the models produced and susceptibility to errors [70]. In this respect, any
597 machine learning procedure is susceptible to overfitting, although this depends upon context;
598 models are often intended to be generalisable to a wide range of datasets beyond the dataset
599 that they were trained on. Arguably, this would be less of an issue for the current study because,
600 rather than producing a tool to predict health outcomes of other dogs, our aim was to identify
601 the relative strength of the different predictors within the dataset itself. Further, many variables
602 of significance were identified using both multiple logistic regression and machine learning
603 predictive modelling, whilst their relative contribution to model prediction were broadly
604 similar (see below). Therefore, despite the limitations of the different modelling techniques,
605 our findings should be valid.

606 Using logistic regression, the main variables associated with dog health in the best-fit
607 models were the age of the dog, the number of times they had visited their veterinary practice
608 within the last 12 months, and whether the dog was receiving medication; in some models,
609 whether the dog had been switched to a therapeutic food was also of significance, especially
610 where data from dogs were described as "generally healthy, but with minor or infrequent
611 problems" by their owners (e.g., ordinal logistic regression and binary logistic regression on
612 the *any health problem* variable). Depending on the model used, these variables explained
613 between 35% and 58% of variance within the dataset, with very little additional variance (3%

614 to 6%, depending on the model) gained by the addition of any or all other predictor variables.
615 Many of the same variables were also identified using XGBoost, but other predictor variables
616 were also identified including owner age and owner education. However, in contrast to the
617 findings of the previous study [15], the diet of the dog had little to no effect in any of the final
618 models. Therefore, we could not find convincing evidence that any particular food type
619 (conventional, vegan, vegetarian or raw) has any clinically meaningful positive or negative
620 association with dog health status. As with the original study [15], these analyses were limited
621 by the dataset used, including its cross-sectional design, meaning that causality of associations
622 cannot be assumed, and its reliance on subjective owner opinions of health status, which might
623 not reflect the actual health of the dogs. Future studies should be considered, for example
624 cohort studies or randomised trials utilising objective measures of health such as information
625 from electronic patient record as or formal diagnoses of disease by a veterinarian.

626 A key limitation of using questionnaires to ascertain the opinions of owners about dog
627 health is that such information might be affected by owner recollection leading to errors in
628 classification of disease status, so-called response bias [71]. Responses might be biased for
629 various reasons including acquiescence, courtesy, the order of questions or even the perceived
630 social desirability of the responses [71]. One form of response bias is recall bias whereby the
631 recollections of the owner are affected by their past experiences [72]. Owners of dogs with
632 health problems might search their memories more thoroughly for possible risk factors (such
633 as diet choice, number of visits to the vet, use of medication etc, diet changes) than owners of
634 dogs that do not have health problems, with the potential to exaggerate any associations
635 identified.

636 It is also feasible that the attitudes and beliefs of owners might either consciously or
637 unconsciously have influenced responses about the health of their dog. For example, owners
638 who believe a particular type of diet to be optimal for the health of their dog, might be more

639 likely to perceive their dog to be healthy, whether or not this is actually the case. Whilst such
640 a bias might feasibly affect any diet type, a greater bias might be expected with diets perceived
641 to be 'unconventional'. Although increasing in popularity [1], both vegan diets and raw meat
642 diets are still an uncommon choice for owners, and many veterinary professionals either do not
643 recommend them or might even advise against their use. Current evidence suggests that both
644 such diet types are often not formulated appropriately to be complete and balanced for essential
645 nutrients [5,8,9-11]. Further, there are also concerns with feeding raw food given the potential
646 for contamination either with bacteria of possible pathogenic potential or bacteria resistant to
647 antimicrobials, both of which might pose a risk to the health of owners and other in-contact
648 people [11-13]. Given such health concerns, owners who make an active choice to feed either
649 a raw or vegan diet could be more defensive about this diet choice, compared with owners who
650 feed conventional diets, and this attitude which might unconsciously have biased responses
651 about the health of their dog, for example, minimising the significance of any illnesses. In the
652 original study questionnaire, the authors did attempt to minimise any influence of canine diet
653 variables on reported health status, by placing the health-related questions before most other
654 variables [15]. However, it is unclear as to whether this would adequately eliminate pre-
655 existing unconscious biases resulting from diet choice.

656 One variable that featured consistently in all models was the age of the dog, whereby
657 illness severity worsened as age increased, and was worst of all for the oldest dogs (9 years or
658 older). Perceptions of health in senior dogs are likely to be associated with quality of life,
659 which is known to be negatively associated with age [73]. However, it might also be because
660 many chronic diseases have associations with increasing age, including osteoarthritis [65],
661 obesity [74], diabetes mellitus [67] and many forms of cancer [75]. It is noteworthy that such
662 chronic diseases are also negatively associated with quality of life [76-79]. An alternative
663 explanation for the negative association between age and health in the current study is the fact

664 that owners of older dogs perceive their health to be poorer simply because they expect older
665 dogs to be less healthy [80].

666 A second association identified with owner perceptions of health was the use of
667 medications although, once again, causality cannot be assumed. It is unclear whether such
668 dogs are receiving these medications because they are genuinely less healthy (which is why
669 they were prescribed), whether the owner perceives their dog to be less healthy simply because
670 they are receiving medications, or both. Finally, the association might be indirect, if both the
671 outcome and predictor variable were associated with a third variable. In this regard, there was
672 also a strong positive association between health status and the frequency of veterinary visits,
673 whilst a positive correlation between frequency of veterinary visits and receiving medication
674 was also evident, meaning that all three variables are likely to be inter-related. Further, in some
675 models, switching to a therapeutic diet was also associated with health status. This would be
676 another variable associated with the frequency of veterinary visits. Besides these, several
677 additional associations were identified with XGBoost, but not with logistic regression, the most
678 important of which were owner age and education. Owners that were older and those with a
679 higher level of education perceived their dogs to be healthier than did other owners. Further
680 studies would be required to explore all such associations further and determine their true
681 significance to canine health.

682 Although the study was large, the population studied was not representative of the typical
683 dog-owning public. In this respect, 33% and 13% of owners reportedly fed their dog raw and
684 vegan diets, respectively, which is greater than expected; for example, in a recent UK survey,
685 the proportion of owners feeding raw and vegan diets were 7% and <1%, respectively [1].
686 Further, the proportion of the UK human population reportedly consuming a vegan diet is
687 estimated to be between 2 and 3% [2] whilst, in the current study, 10% were of owners
688 identified as vegan. This was of particular significance given the identified of clustering of

owner-animal metadata based on the owner's diet choice. Besides this, and despite involving responders from many countries, the geographical split was unbalanced with 73% of survey responses coming from the UK. Although the reasons for such imbalances are not clear, it might well have resulted from the method of owner recruitment. In this respect, the study was widely advertised via social media but, given concerns that owners feeding dogs unconventional diets might be under-represented, relevant interest groups were actively targeted and invited to attend. Therefore, the recruitment strategy might have inadvertently generated an unrepresentative study population which, in turn, may have influenced the results obtained, as discussed above.

In addition to the unrepresentative study population, multiple associations amongst groups were identified with the potential for confounding amongst variables. For example, feeding a raw diet was more commonly reported by owners from the UK, whilst feeding a vegan diet was more common in owners from Europe. Further, there were positive associations between owner income and education, and between dog age and neuter status. In logistic regression analyses, we attempted to account for these associations by determining multicollinearity, and refining models where this might be a problem. However, this does not necessarily guarantee that variance from each factor in the models was correctly assigned. As discussed above, this was addressed using XGBoost, and many of the variables identified with logistic regression were also identified with this technique. Other variables of significance (albeit lesser) were also identified including owner age, education and size category. As such analyses are exploratory, the reasons for such associations with owner opinion on dog health require further study.

The main limitations to the current study have already been discussed above, including the fact that owner-reported health was the main outcome measure and the fact that the study population was not representative of the general dog-owning population. Further, significant

714 data pre-processing was required to ensure adequate group sizes for statistical analyses. The
715 disadvantage of such an approach was that many smaller groups needed to be excluded,
716 meaning that we might have missed some variables with a potential impact on owner-reported
717 canine health. A final possible limitation was the fact that we chose to include healthcare
718 variables in addition to owner and dog variables. However, there might also be confounding
719 between these healthcare variables (such as number of veterinary visits, being on medication
720 or use of a therapeutic diet) and dietary choices. For example, owners who are sceptical about
721 conventional diets and, instead choose to feed vegan or uncooked foods, might also be sceptical
722 about either visiting their veterinarian or allowing their dog to receive medication. This
723 possibility is supported by the fact that there were negative associations between the use of
724 medications and feeding either a raw or vegan diet, whilst frequency of veterinary visits was
725 also negatively associated with feeding a raw diet. Therefore, a downside of including the
726 healthcare variables in multiple logistic regression analyses is that their presence might have
727 adversely affected the associations seen with the owner-animal metadata leading to biased
728 models. Arguably, however, the use of XGBoost should offset the limitation of confounding
729 from healthcare variables given that these techniques do not make prior assumptions about data
730 distribution and can take account of complexity in interactions amongst variables [69].
731 However, to ensure that the effect of owner-animal metadata had not been unfairly penalised
732 by the inclusion of the healthcare variables, in our final machine learning predictive modelling,
733 we examined models containing only the owner-animal metadata and not the healthcare
734 variables. In these models, dog age remained as the variable of most importance, but other
735 variables of lesser importance were identified that had not been seen on the original models
736 (including UK location, dog sex and dog size), whilst the limited effect of diet remained at a
737 similar level (i.e., use of a vegan diet was only the 7th most important predictor variable).
738 Further, overall, the ROC curve highlighted that model prediction was unacceptably poor. This

suggests that besides dog age, the rest of the owner-animal metadata are poor predictors of owner-perceived dog health.

Conclusions

Variables most strongly associated with owner perceptions of health are age of dog and the healthcare variables including number of visits to a veterinarian, administration of medications and use of a therapeutic diet. These results conflict with previously-published research [15] and suggest that, when other variables are accounted for, diet-related variables have minimal, if any, genuine association with owner-perceived canine health.

Acknowledgements

The authors wish to acknowledge the authors of the original paper [15] for providing making their data freely available to other researchers for further analysis.

Funding statement

This study was not funded by any research grant. The costs of open access publishing were covered by the University of Liverpool Institutional Open Access Fund (https://www.liverpool.ac.uk/open-research/open-access/open-access-funding/).

Conflict of interest statement

AJG is an employee of the University of Liverpool, but his position is financially-supported by Royal Canin. AJG has also received financial remuneration and gifts for providing educational material, speaking at conferences, and consultancy work, all unrelated

762 to the current study. Royal Canin had no involvement in any aspect of the current work
763 including study design, data analysis, drafting the manuscript or the decision to submit the
764 work for publication.
765 RBJ is an employee of the University of Liverpool, whose salary is financially-supported
766 by the Higher Education Funding Council for England. RBJ also receives a stipend from
767 United Kingdom Research and Innovation for work as a panellist for the Biotechnology and
768 Biological Sciences Research Council.
769

770 # Research data for this article

771 The original research data are accessible at https://rebrand.ly/2020/pet-food-consumers-
772 study. All statistical reports and codes are available in the supplementary information (S1-S5
773 Files) and online (https://github.com/RichardBJ/CanineHealth22).
774

775 # References

776 1. PDSA. PAW Report 2022. [cited 2022 Dec 19]. Available from:
777 https://www.pdsa.org.uk/media/12965/pdsa-paw-report-2022.pdf.
778 2. YouGov 2022. Dietary choices of Brits (e.g. vegetarian, flexitarian, meat-eater etc).
779 2022 [cited 2022 Oct 20]. Available from:
780 https://yougov.co.uk/topics/lifestyle/trackers/dietery-choices-of-brits-eg-
781 vegeterian-flexitarian-meat-eater-etc.
782 3. New Nutrition Business. 10 key trends in food, nutrition & health 2020. 2022 [cited
783 2022 Dec 19]. Available from: https://www.new-nutrition.com/keytrend?id=135.
784 4. Gallup. Snapshot: few Americans vegetarian or vegan [internet]. 2018 [cited 2022
785 Oct 20]. Available from: https://news.gallup.com/poll/238328/snapshot-few-

786 americans-vegetarian-
787 vegan.aspx?g_source=link_NEWSV9&g_medium=NEWSFEED&g_campaign=it
788 em_&g_content=Snapshot%3a%2520Few%2520Americans%2520Vegetarian%25
789 20or%2520Vegan.
790 5. Dodd SAS, Adolphe JL, Verbrugghe A. Plant-based diets for dogs. J Am Vet Med
791 Assoc. 2018 Dec 1;253(11):1425-1432. doi: 10.2460/javma.253.11.1425
792 6. Dodd S, Khosa D, Dewey C, Verbrugghe A. Owner perception of health of North
793 American dogs fed meat- or plant-based diets. Res Vet Sci. 2022 Dec;149:36-46.
794 doi: 10.1016/j.rvsc.2022.06.002
795 7. Morelli G, Bastianello S, Catellani P, Ricci R. Raw meat-based diets for dogs:
796 survey of owners' motivations, attitudes and practices. BMC Vet Res. 2019 Mar
797 4;15(1):74. doi: 10.1186/s12917-019-1824-x
798 8. Dodd SAS, Shoveller AK, Fascetti AJ, Yu ZZ, Ma DWL, Verbrugghe A. A
799 Comparison of Key Essential Nutrients in Commercial Plant-Based Pet Foods Sold
800 in Canada to American and European Canine and Feline Dietary Recommendations.
801 Animals (Basel). 2021 Aug 9;11(8):2348. doi: 10.3390/ani11082348
802 9. Larsen JA, Parks EM, Heinze CR, Fascetti AJ. Evaluation of recipes for home-
803 prepared diets for dogs and cats with chronic kidney disease. J Am Vet Med Assoc.
804 2012 Mar 1;240(5):532-8. doi: 10.2460/javma.240.5.532
805 10. Stockman J, Fascetti AJ, Kass PH, Larsen JA. Evaluation of recipes of home-
806 prepared maintenance diets for dogs. J Am Vet Med Assoc. 2013 Jun
807 1;242(11):1500-5. doi: 10.2460/javma.242.11.1500. Erratum in: J Am Vet Med
808 Assoc. 2014 Jul 15;245(2):177

809 11. Vecchiato CG, Schwaiger K, Biagi G, Dobenecker B. From nutritional adequacy to
810 hygiene quality: a detailed assessment of commercial raw pet-food for dogs and
811 cats. Animals (Basel). 2022 Sep 13;12(18):2395. doi: 10.3390/ani12182395
812 12. Groat EF, Williams NJ, Pinchbeck G, Warner B, Simpson A, Schmidt VM. UK dogs
813 eating raw meat diets have higher risk of Salmonella and antimicrobial-resistant
814 Escherichia coli faecal carriage. J Small Anim Pract. 2022 Jun;63(6):435-441. doi:
815 10.1111/jsap.13488
816 13. Sealey JE, Hammond A, Mounsey O, Gould VC, Reyher KK, Avison MB.
817 Molecular ecology and risk factors for third-generation cephalosporin-resistant
818 Escherichia coli carriage by dogs living in urban and nearby rural settings. J
819 Antimicrob Chemother. 2022 Aug 25;77(9):2399-2405. doi: 10.1093/jac/dkac208
820 14. Anturaniemi J, Barrouin-Melo SM, Zaldivar-López S, Sinkko H, Hielm-Björkman
821 A. Owners' perception of acquiring infections through raw pet food: a
822 comprehensive internet-based survey. Vet Rec. 2019 Nov 30;185(21):658. doi:
823 10.1136/vr.105122
824 15. Knight A, Huang E, Rai N, Brown H. Vegan versus meat-based dog food: Guardian-
825 reported indicators of health. PLoS One. 2022 Apr 13;17(4):e0265662. doi:
826 10.1371/journal.pone.0265662.
827 16. Chen, T., Guestrin, C. XGBoost: a scalable tree boosting system [cited 2022 Nov
828 8] Available from: https://arxiv.org/abs/1603.02754#.
829 17. R Core Team. R: A language and environment for statistical computing. R
830 Foundation for Statistical Computing, Vienna, Austria. 2022 [cited 2022 Nov 30].
831 Available from: https://www.R-project.org/

18. Lesnoff, M., Lancelot, R. (2012). aod: analysis of overdispersed data. R package version 1.3.2. 2012 [cited 2022 Dec 19]. Available from: http://cran.r-project.org/package=aod

19. Robinson D, Hayes A, Couch S. broom: convert statistical objects into tidy tibbles; R package version 0.8.0. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=broom.

20. Fox J, Weisberg S. An R Companion to Applied Regression, Third Edition; 2019 [cited 2022 Dec 19]. Available from: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

21. Kuhn M. caret: classification and regression training; R package version 6.0-93. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=caret.

22. Wei T, Simko V (2021). R package 'corrplot': visualization of a correlation matrix (version 0.92). 2021 [cited 2022 Nov 8]. Available from: https://github.com/taiyun/corrplot.

23. Dowle M, Srinivasan A. data.table: extension of `data.frame`; R package version 1.14.6. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=data.table.

24. Hahsler M, Piekenbrock M (2022). dbscan: density-based spatial clustering of applications with noise (DBSCAN) and related algorithms; R package version 1.1-11. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=dbscan.

25. Signorell A, Aho K, Alfons A, Anderegg N,Aragon T, Arachchige A, et al. DescTools: tools for descriptive statistics; R package version 0.99.47. 2022 [cited 2022 Dec 19]. Available from: https://cran.r-project.org/package=DescTools

26. Wickham H, Hester J, Chang W, Bryan J. devtools: tools to make developing R packages easier; R package version 2.4.3. 2021 [cited 2022 Dec 19]. Available: https://CRAN.R-project.org/package=devtools.

27. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.9. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=dplyr.

28. Nakazawa M. fmsb: functions for medical statistics book with some demographic data. R package version 0.7.3. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=fmsb

29. R Core Team. foreign: read data stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ..; R package version 0.8-82. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=foreign.

30. Jay M. generalhoslem: goodness of fit tests for logistic regression models. R package version 1.3.4. 2019 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=generalhoslem

31. Wickham H. ggplot2: elegant graphics for data analysis. 3rd ed. New York: Springer-Verlag New York, 2016.

32. Arnold J. ggthemes: extra themes, scales and geoms for 'ggplot2'; R package version 4.2.4. 2021 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=ggthemes.

33. Vanegas L, Rondón L, Paula G. (2022). _glmtoolbox: set of tools to data analysis using generalized linear models; R package version 0.1.4. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=glmtoolbox.

880 34. Ugba E. _gofcat: goodness-of-fit measures for categorical response models; R
881 package version 0.1.2. 2022 [cited 2022 Dec 19]. Available from:
882 https://CRAN.R-project.org/package=gofcat.
883 35. Galili T, O'Callaghan A, Sidi J, Sievert C. heatmaply: an R package for creating
884 interactive cluster heatmaps for online publishing. Bioinformatics. 2018 May
885 1;34(9):1600-1602. doi: 10.1093/bioinformatics/btx657
886 36. Harrell Jr F. Hmisc: Harrell miscellaneous. R package version 4.7-0 [internet]. 2022
887 [cited 2022 Dec 12]. Available from: https://CRAN.R-project.org/package=Hmisc
888 37. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting
889 increases detection power in genome-scale multiple testing. Nat Methods. 2016
890 Jul;13(7):577-80. doi: 10.1038/nmeth.3885
891 38. Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic
892 regression models in large samples: A modification of the Hosmer-Lemeshow test.
893 Biometrics. 2020 Jun;76(2):549-560. doi: 10.1111/biom.13249
894 39. Venables W N, Ripley B D. Modern Applied Statistics with S. 4th ed. New York:
895 Springer; (2002. ISBN 0-387-95457-0
896 40. Gorman B. mltools: machine learning tools. R package version 0.3.5, 2018 [cited
897 2022 Dec 12]. Available from: https://CRAN.R-project.org/package=mltools.
898 41. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an
899 open-source package for R and S+ to analyze and compare ROC curves. BMC
900 Bioinformatics. 2011 Mar 17;12:77. doi: 10.1186/1471-2105-12-77
901 42. Wickham H, Bryan J. readxl: read Excel files; R package version 1.4.0, 2022 [cited
902 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=readxl.

43. Silge J, Chow F, Kuhn M, Wickham H. rsample: General Resampling Infrastructure; R package version 1.1.0, 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=rsample.

44. Wickham H. Reshaping data with the reshape package. J Stat Soft 2007;21(12):1-20. doi: 10.18637/jss.v021.i12

45. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005 Oct 15;21(20):3940-1. doi: 10.1093/bioinformatics/bti623

46. Kassambara A. rstatix: pipe-friendly framework for basic statistical tests. R package version 0.7.0. 2021 [cited 2022 Nov 8]. Available from: https://CRAN.R-project.org/package=rstatix.

47. Qiu Y, details. aotisSfAf. showtext: using fonts more easily in R graphs; R package version 0.9-5. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=showtext.

48. Siriseriwan W. smotefamily: A collection of oversampling techniques for class imbalance problem based on SMOTE; R package version 1.3.1. 2019 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=smotefamily.

49. Wickham H. stringr: simple, consistent wrappers for common string pperations; R package version 1.4. 2019 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=stringr.

50. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, Francois R, et al. Welcome to the tidyverse. J Open Source Software. 2019;4(43):1686. doi: 10.21105/joss.01686

51. Konopka T. umap: uniform manifold approximation and projection. R package version 0.2.9.0. 2022 [cited 2022 Dec 5]. Available from: https://CRAN.R-project.org/package=umap.

52. Yee TW. VGAM: vector generalized linear and additive models; R package version 1.1-6. 2022 [cited 2022 Dec 19]. Available from: https://CRAN.R-project.org/package=VGAM.

53. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426v3. 2018 [cited 2022 Dec 19]. doi: 10.48550/arXiv.1802.03426

54. Kendall MG. The treatment of ties in ranking problems. Biometrika. 1945 Nov;33:239-51. doi: 10.1093/biomet/33.3.239

55. Friendly M. Corrgrams: exploratory displays for correlation matrices. Am Stat. 2002;56:316–324. doi: 10.1198/000313002533

56. Murtagh, F., Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?. J Classif. 2014;31:274–295. doi: 10.1007/s00357-014-9161-z

57. Cohen J. A power primer. Psychol Bull. 1992 Jul;112(1):155-159. doi: 10.1037//0033-2909.112.1.155

58. Schwartz, G. Estimating the dimension of a model. Ann Stat. 1978;6:461-464.

59. Neath, A.A., Cavanaugh, J.E. The Bayesian information criterion: background, derivation, and applications. WIREs Comput Stat 2012;4:199-203. doi: 10.1002/wics.199

60. Fox J, Monette G. Generalized Collinearity Diagnostics. J Am Stat Assoc. 1992;87:417, 178-183, DOI: 10.1080/01621459.1992.10475190

61. Ugba, E. R. and Gertheiss, J. (2018). An Augmented Likelihood Ratio Index for Categorical Response Models. In Proceedings of 33rd International Workshop on Statistical Modelling' 2018 July 15-20; Bristol UK. p. 293-298.

62. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78:691-692. doi: 10.1093/biomet/78.3.691

63. Wong SC, Gatt A, Stamatescu V, McDonnell MD. Understanding data augmentation for classification: when to warp? 2016 [cited 22 Nov 8]. Available from: https://arxiv.org/abs/1609.08764.

64. Westgarth C, Pinchbeck GL, Bradshaw JW, Dawson S, Gaskell RM, Christley RM. Factors associated with dog ownership and contact with dogs in a UK community. BMC Vet Res. 2007 Apr 3;3:5. doi: 10.1186/1746-6148-3-5

65. Anderson KL, O'Neill DG, Brodbelt DC, Church DB, Meeson RL, Sargan D, et al. Prevalence, duration and risk factors for appendicular osteoarthritis in a UK dog population under primary veterinary care. Sci Rep. 2018 Apr 4;8(1):5641. doi: 10.1038/s41598-018-23940-z

66. Zapata I, Moraes LE, Fiala EM, Zaldivar-Lopez S, Couto CG, Rowell JL, et al. Risk-modeling of dog osteosarcoma genome scans shows individuals with Mendelian-level polygenic risk are common. BMC Genomics. 2019 Mar 19;20(1):226. doi: 10.1186/s12864-019-5531-6

67. Heeley AM, O'Neill DG, Davison LJ, Church DB, Corless EK, Brodbelt D. Diabetes mellitus in dogs attending UK primary-care practices: frequency, risk factors and survival. Canine Genet Epidemiol. 2020;7: 6. doi: 10.1186/s40575-020-00087-7

68. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011 Oct;18(10):1099-104. doi: 10.1111/j.1553-2712.2011.01185.x

975 69. Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B,
976 Steyerberg EW; CENTER-TBI collaborators. Machine learning algorithms
977 performed no better than regression models for prognostication in traumatic brain
978 injury. J Clin Epidemiol. 2020 Jun;122:95-107. doi: 10.1016/j.jclinepi.2020.03.005
979 70. Data-Flair. Advantages and disadvantages of machine learning language [internet].
980 [cited 2022 Dec 9]. Available from: https://data-flair.training/blogs/advantages-
981 and-disadvantages-of-machine-learning/.
982 71. Villar A. Response Bias. Encyclopedia of Survey Research Methods. Lavrakas PJ,
983 editor. Thousand Oaks, Sage Publications Ltd. p. 752-753.
984 72. Grimes DA, Schulz KF. Bias and causal associations in observational research.
985 Lancet. 2002 Jan 19;359(9302):248-52. doi: 10.1016/S0140-6736(02)07451-2
986 73. Rodger S, Scott EM, Nolan A, Wright AK, Reid J. Effect of Age, Breed, and Sex
987 on the Health-Related Quality of Life of Owner Assessed Healthy Dogs. Front Vet
988 Sci. 2021 Feb 5;8:603139. doi: 10.3389/fvets.2021.603139
989 74. Lund EM, Armstrong PA, Kirk CA, Klausner JS. Prevalence and risk factors for
990 obesity in adult dogs from private US veterinary practices. Intern J Appl Res Vet
991 Med. 2006;4[2]:177-186.
992 75. Merlo DF, Rossi L, Pellegrino C, Ceppi M, Cardellino U, Capurro C, et al. Cancer
993 incidence in pet dogs: findings of the Animal Tumor Registry of Genoa, Italy. J Vet
994 Intern Med. 2008 Jul-Aug;22(4):976-84. doi: 10.1111/j.1939-1676.2008.0133.x
995 76. German AJ, Holden SL, Wiseman-Orr ML, Reid J, Nolan AM, Biourge V, Morris
996 PJ, Scott EM. Quality of life is reduced in obese dogs but improves after successful
997 weight loss. Vet J. 2012 Jun;192(3):428-34. doi: 10.1016/j.tvjl.2011.09.015

998 77. Niessen SJ, Powney S, Guitian J, Niessen AP, Pion PD, Shaw JA, et al. Evaluation
999 of a quality-of-life tool for dogs with diabetes mellitus. J Vet Intern Med. 2012 Jul-
1000 Aug;26(4):953-61. doi: 10.1111/j.1939-1676.2012.00947.x
1001 78. Iliopoulou MA, Kitchell BE, Yuzbasiyan-Gurkan V. Development of a survey
1002 instrument to assess health-related quality of life in small animal cancer patients
1003 treated with chemotherapy. J Am Vet Med Assoc. 2013 Jun 15;242(12):1679-87.
1004 doi: 10.2460/javma.242.12.1679
1005 79. Roberts C, Armson B, Bartram D, Belshaw Z, Capon H, Cherry G, et al.
1006 Construction of a conceptual framework for assessment of health-related quality of
1007 life in dogs with osteoarthritis. Front Vet Sci. 2021 Sep 24;8:741864. doi:
1008 10.3389/fvets.2021.741864
1009 80. Wallis LJ, Radford AD, German AJ, Jackson J, Belshaw Z, Kubinyi E, et al. Just
1010 old age" - a qualitative investigation of UK owner experiences and attitudes to
1011 ageing of senior and geriatric dogs. Proceedings of the 30th International Society
1012 for Anthrozoology Conference; 2021 June 22-24; online. p81.
1013
1014

Supplementary material

S1 File. Statistical report containing details of all data pre-processing steps and initial data visualisation using UMAP.

S2 File. Statistical report containing details of all statistical analyses for the Kendall's rank correlation and also XGBoost modelling for the 'any health problem' variables.

S3 File. Statistical report containing details of all ordinal logistic regression modelling for the 'illness severity' variable.

S4 File. Statistical report containing details of all binary logistic regression modelling for the 'any health problem' variable.

S5 File. Statistical report containing details of all binary logistic regression modelling for the 'significant illness' variable.

S6 File. Statistical report containing details of XGBoost modelling for the 'significant illness' variable.

S1 Table. Kendall's tau correlation coefficients for the naïve associations amongst owner-animal metadata variables in Fig 2.

S2 Table. Respective log *P*-values for the Kendall's tau correlations amongst owner-animal metadata variables in Fig 2 and S1 Table.

S3 Table. Final multiple ordinal logistic regression model examining associations between owner, animal and veterinary variables and illness severity as reported by the owner.

S4 Table. Final multiple binary logistic regression model examining associations between owner, animal and veterinary variables and the presence of any illness as reported by the owner.

S5 Table. Final multiple binary logistic regression model examining associations between owner, animal and veterinary variables and the presence of significant or serious illness as reported by the owner.

S1 Fig. Owner-dog metadata visualised by Uniform Manifold Approximation and Projection with Density-Based Spatial Clustering. Owner-pet metadata were pre-processed with all factors as numeric and subject to dimension reduction with the UMAP projection technique. Owner variables included in this visualisation were *diet*, *sex*, *location*, *education* and *income*; animal variables included were *age*, *sex*, *neuter status*, *diet* and *breed size category*. Healthcare variables were not included. Each individual row of the data (owner-dog combination) contributes one UMAP x and y coordinate, analogous to PC1 and PC2 in a principal component analysis. Four or five distinct clusters were evident. From top to bottom, points are colour-coded by *owner gender*, *income*, *education* and *location* as indicated in the legend.

1064 **S2 Fig. Owner-dog metadata visualised by Uniform Manifold Approximation and**
1065 **Projection with Density-Based Spatial Clustering.** Owner-pet metadata were pre-processed
1066 with all factors as numeric and subject to dimension reduction with the UMAP projection
1067 technique. Owner variables included in this visualisation were *diet*, *sex*, *location*, *education*
1068 and *income*; animal variables included were *age*, *sex*, *neuter status*, *diet* and *breed size category*.
1069 Healthcare variables were not included. Each individual row of the data (owner-dog
1070 combination) contributes one UMAP x and y coordinate, analogous to PC1 and PC2 in a
1071 principal component analysis. Four or five distinct clusters were evident. From top to bottom,
1072 points are colour-coded by *dog age*, *owner age*, *neuter status*, *breed size category* and *dog diet*
1073 as indicated in the legend.
1074
1075 **S3 Fig. Multiple partial proportional odds logistic regression model with *illness severity***
1076 **as the outcome variable and including all owner, animal and healthcare variables.**
1077 The dots represent the odds ratio for each variable, whilst the bars represent 99% confidence
1078 intervals (99%-CI). Variables where the 99%-CI range does not include1.0 (vertical dotted
1079 line) are depicted in red, whilst those that include 1.0 are depicted in blue. Note the logarithmic
1080 scale for the X-axis. Medication and visit number did not meet the partial proportional odds
1081 assumption and, therefore, separate odds ratios and 99% confidence intervals are reported for
1082 the different levels of illness severity ('minor' and 'severe').
1083
1084 **S4 Fig. Multiple binary logistic regression model with *any health problem* as the outcome**
1085 **variable and including all owner, animal and healthcare variables.**
1086 The dots represent the odds ratio for each variable, whilst the bars represent 99% confidence
1087 intervals (99%-CI). Variables where the 99%-CI range does not include1.0 (vertical dotted

line) are depicted in red, whilst those that include 1.0 are depicted in blue. Note the logarithmic scale for the X-axis.

S5 Fig. Multiple binary logistic regression model with *significant illness* as the outcome variable and including all owner, animal and healthcare variables.

The dots represent the odds ratio for each variable, whilst the bars represent 99% confidence intervals (99%-CI). Variables where the 99%-CI range does not include1.0 (vertical dotted line) are depicted in red, whilst those that include 1.0 are depicted in blue. Note the logarithmic scale for the X-axis.

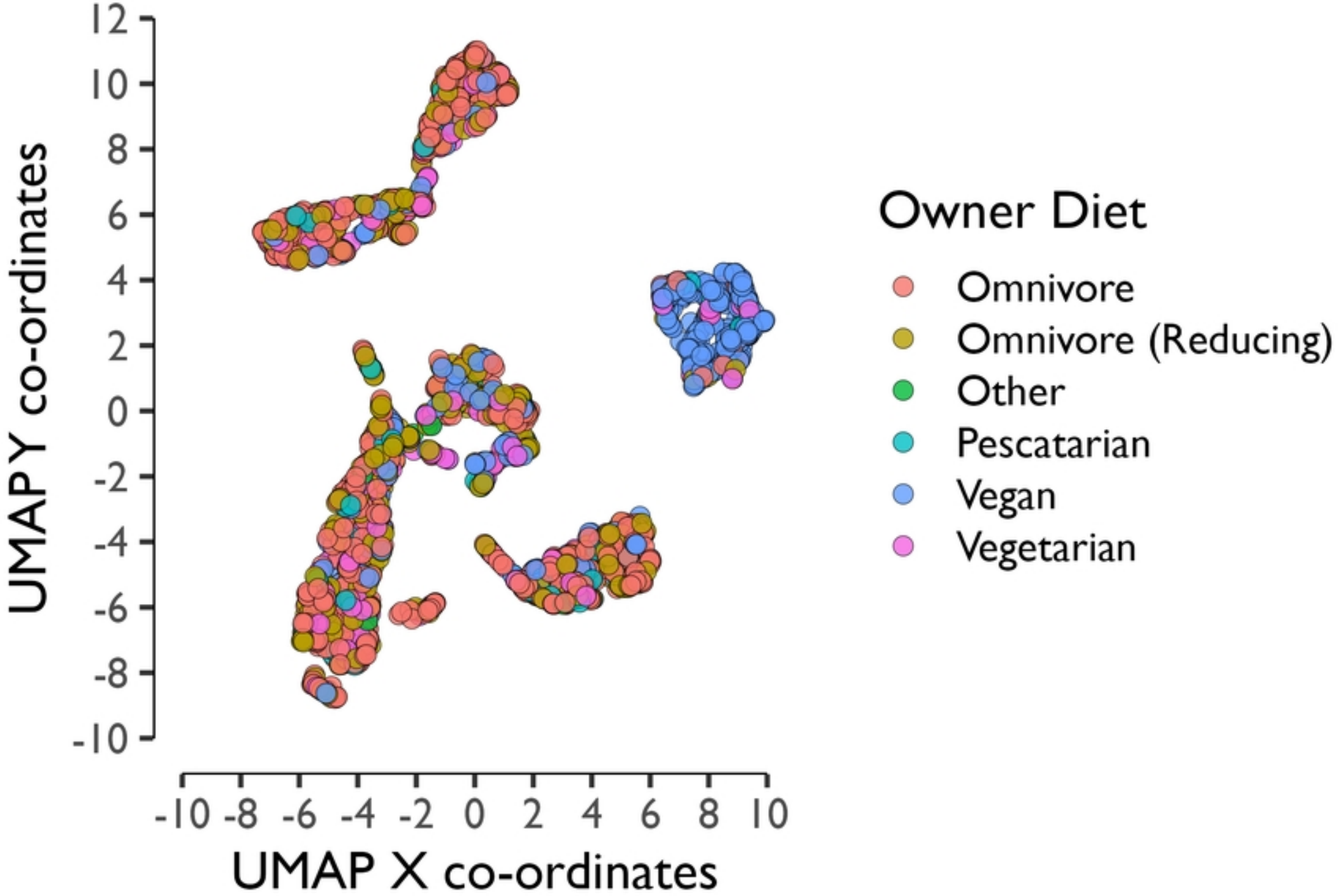
UMAPY co-ordinates
12
10
8
6
4
2
0
-2
-4
-6
-8
-10
-10 -8 -6 -4 -2 0 2 4 6 8 10
UMAP X co-ordinates
Owner Diet
Omnivore
Omnivore (Reducing)
Other
Pescatarian
Vegan
Vegetarian

Figure 1a

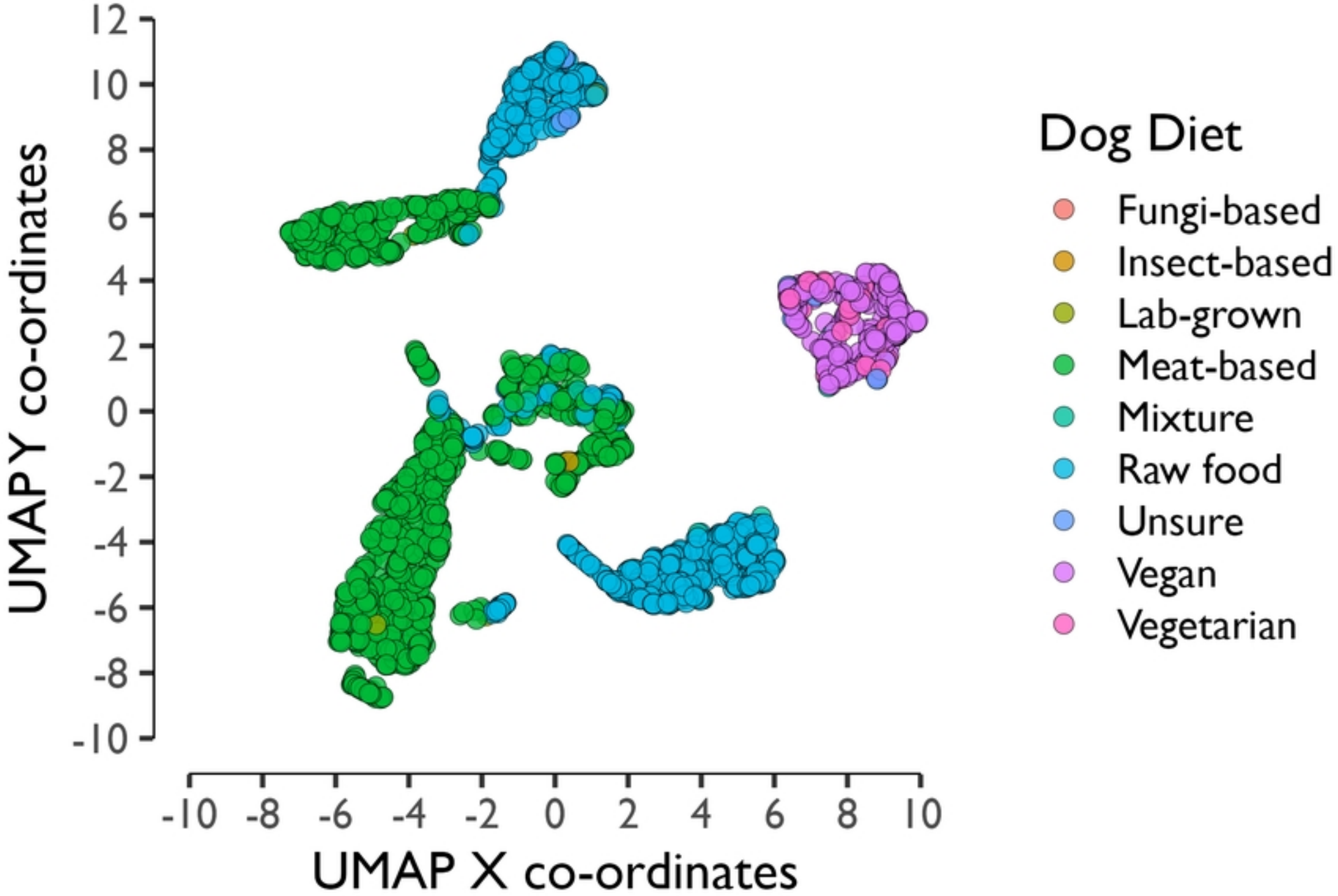
Dog Diet
Fungi-based
Insect-based
Lab-grown
Meat-based
Mixture
Raw food
Unsure
Vegan
Vegetarian
UMAP Y co-ordinates
UMAP X co-ordinates
12
10
8
6
4
2
0
-2
-4
-6
-8
-10
-10
-8
-6
-4
-2
0
2
4
6
8
10

Figure 1b

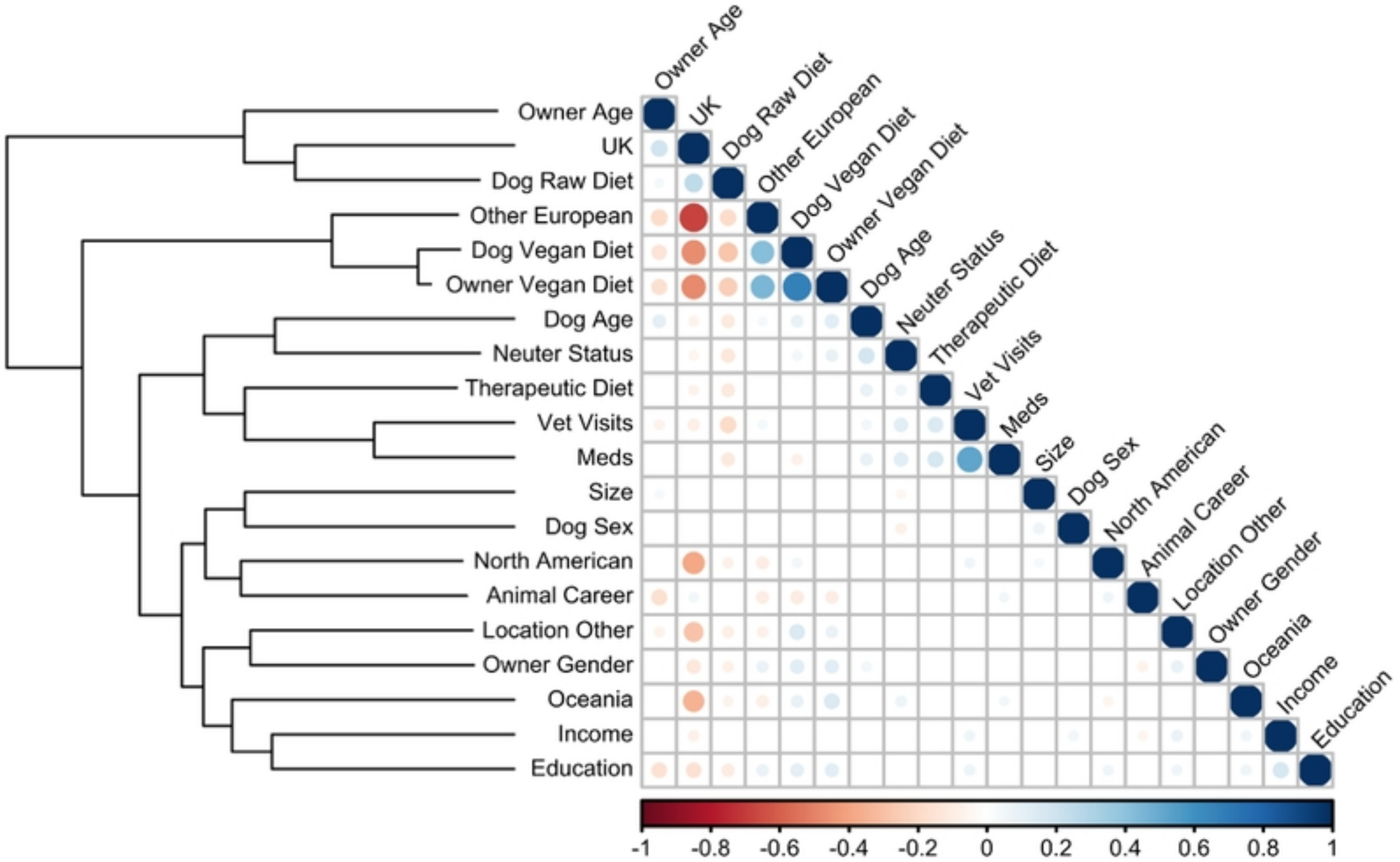
Owner Age
UK
Dog Raw Diet
Other European
Dog Vegan Diet
Owner Vegan Diet
Dog Age
Neuter Status
Therapeutic Diet
Vet Visits
Meds
Size
Dog Sex
North American
Animal Career
Location Other
Owner Gender
Oceania
Income
Education
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Figure 2

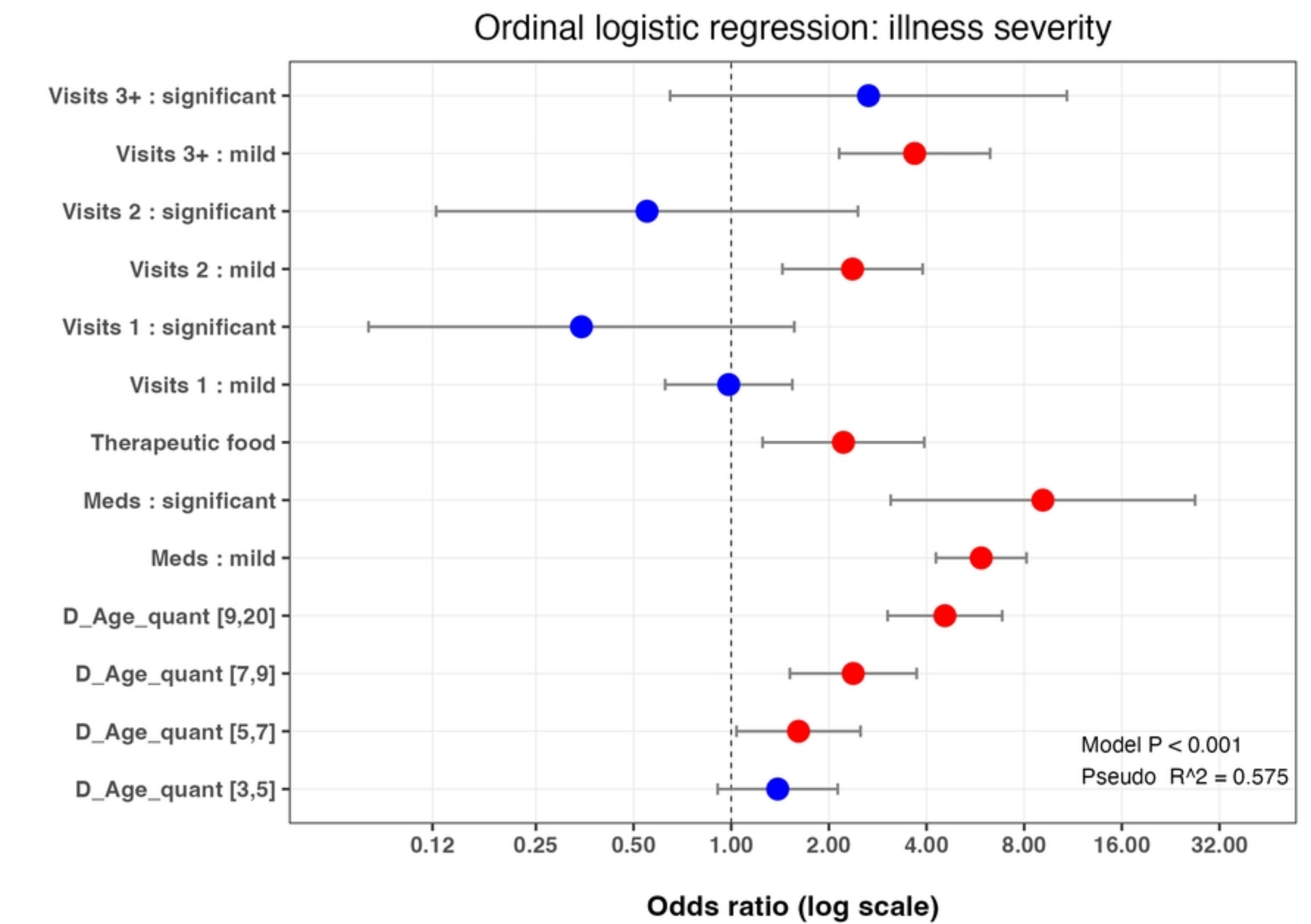
Ordinal logistic regression: illness severity
Visits 3+ : significant
Visits 3+ : mild
Visits 2 : significant
Visits 2 : mild
Visits 1 : significant
Visits 1 : mild
Therapeutic food
Meds : significant
Meds : mild
D_Age_quant [9,20]
D_Age_quant [7,9]
D_Age_quant [5,7]
D_Age_quant [3,5]
Model P < 0.001
Pseudo R^2 = 0.575
0.12
0.25
0.50
1.00
2.00
4.00
8.00
16.00
32.00
Odds ratio (log scale)

Figure 3

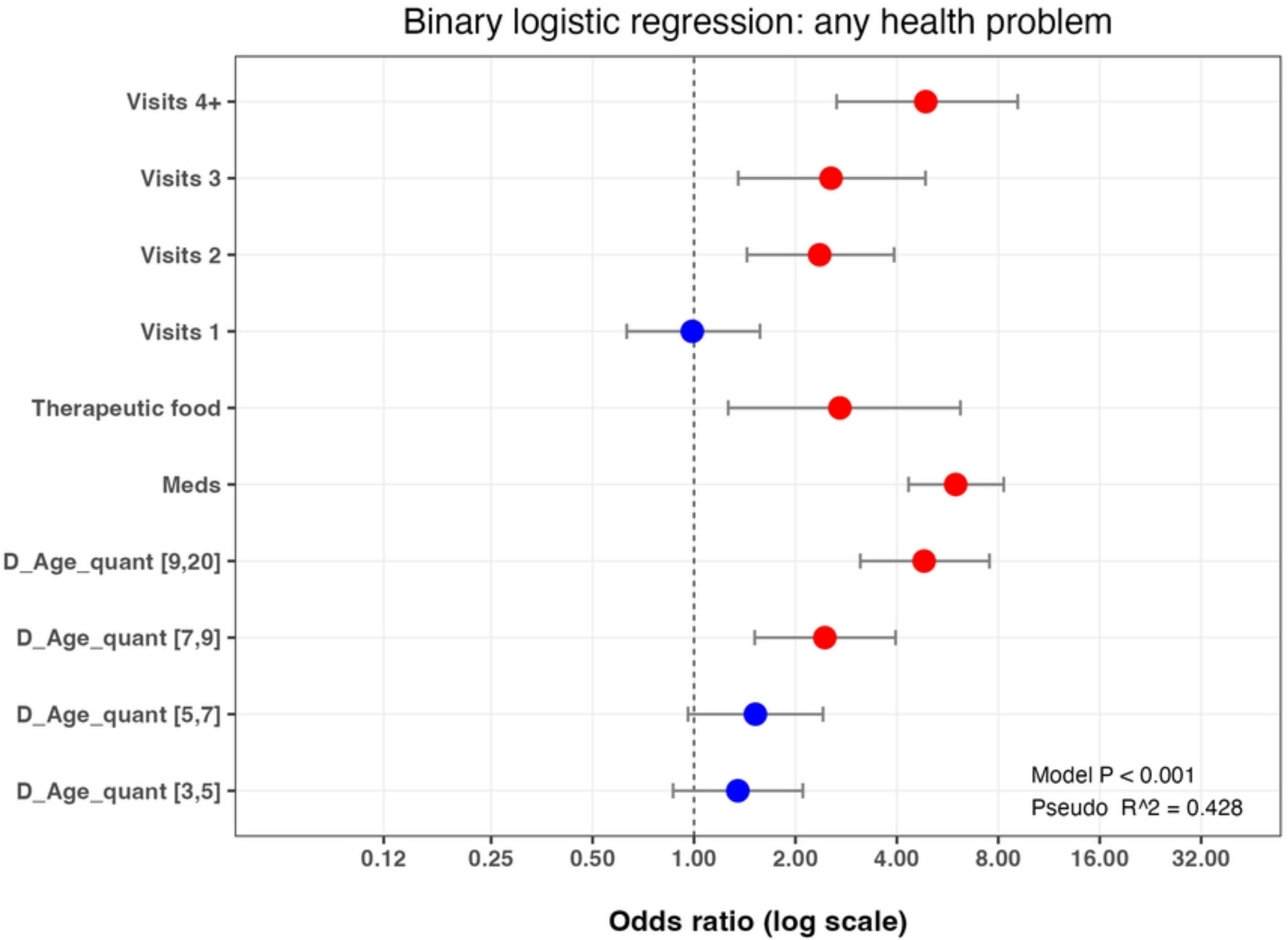
Binary logistic regression: any health problem
Visits 4+
Visits 3
Visits 2
Visits 1
Therapeutic food
Meds
D_Age_quant [9,20]
D_Age_quant [7,9]
D_Age_quant [5,7]
D_Age_quant [3,5]
Model P < 0.001
Pseudo R^2 = 0.428
0.12
0.25
0.50
1.00
2.00
4.00
8.00
16.00
32.00
Odds ratio (log scale)

Figure 4

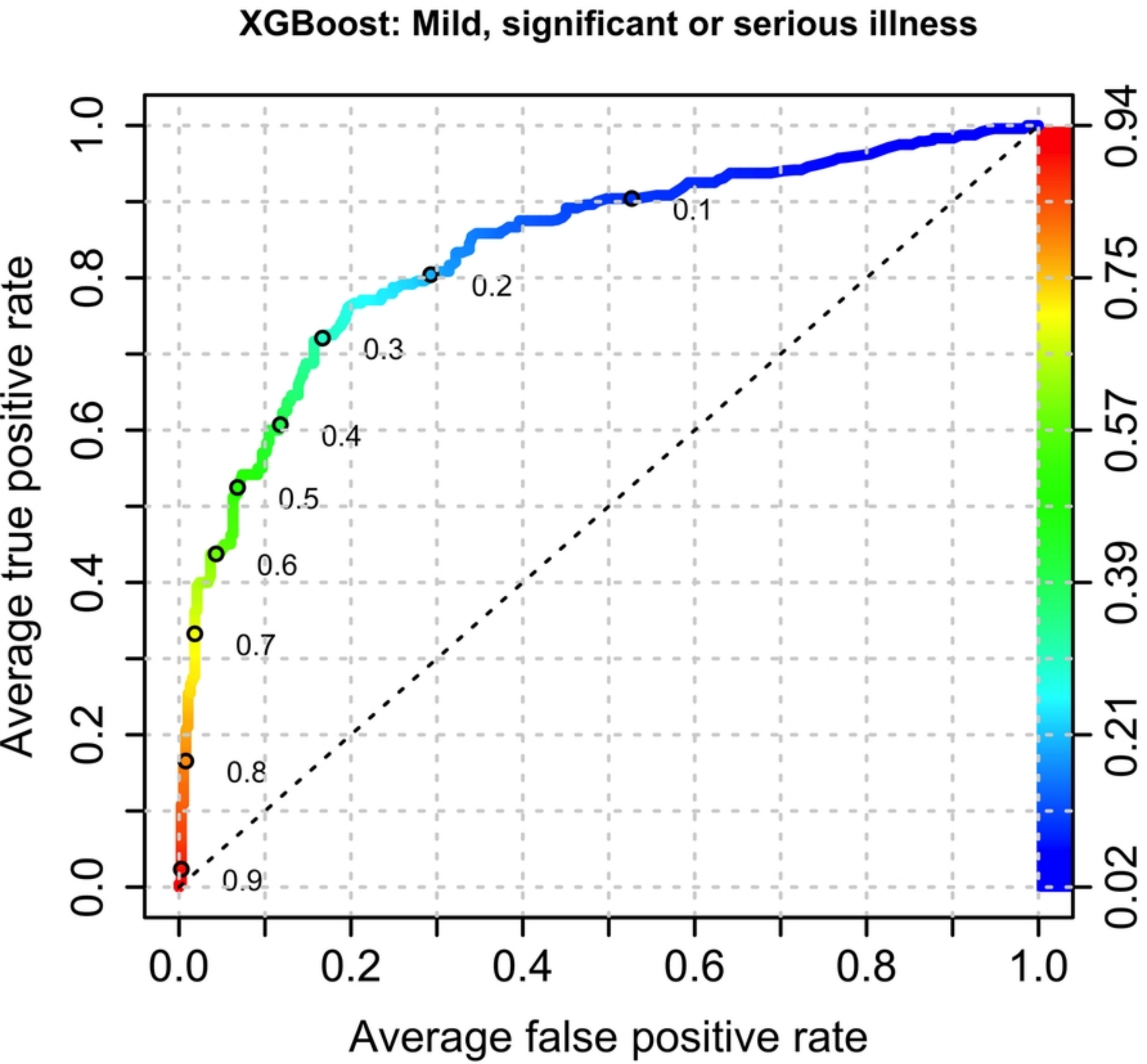
XGBoost: Mild, significant or serious illness
Average true positive rate
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.02
0.21
0.39
0.57
0.75
0.94

Figure 5a

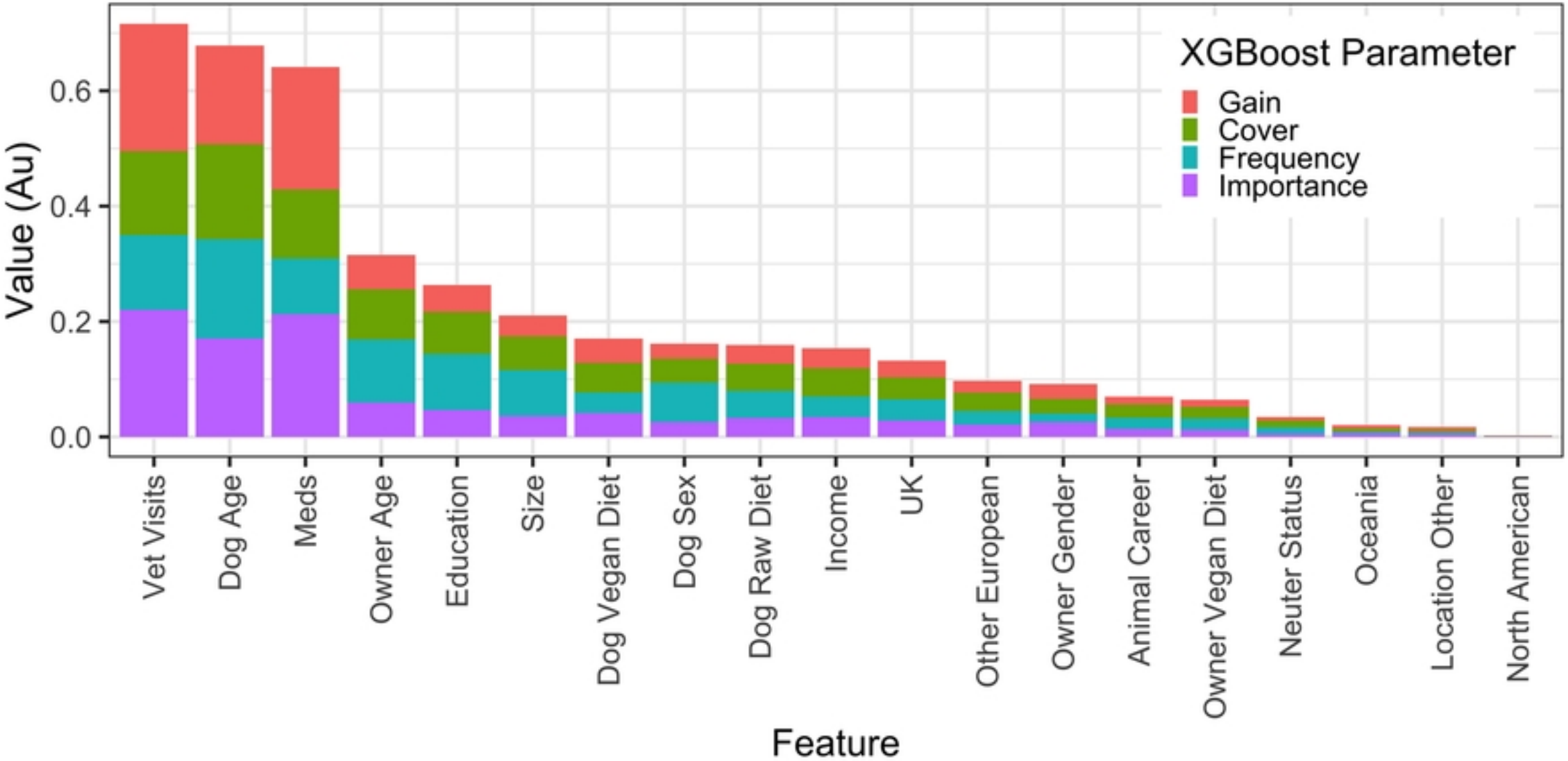
XGBoost Parameter
Gain
Cover
Frequency
Importance
Value (Au)
0.0
0.2
0.4
0.6
Vet Visits
Dog Age
Meds
Owner Age
Education
Size
Dog Vegan Diet
Dog Sex
Dog Raw Diet
Income
UK
Other European
Owner Gender
Animal Career
Owner Vegan Diet
Neuter Status
Oceania
Location Other
North American
Feature

Figure 5b

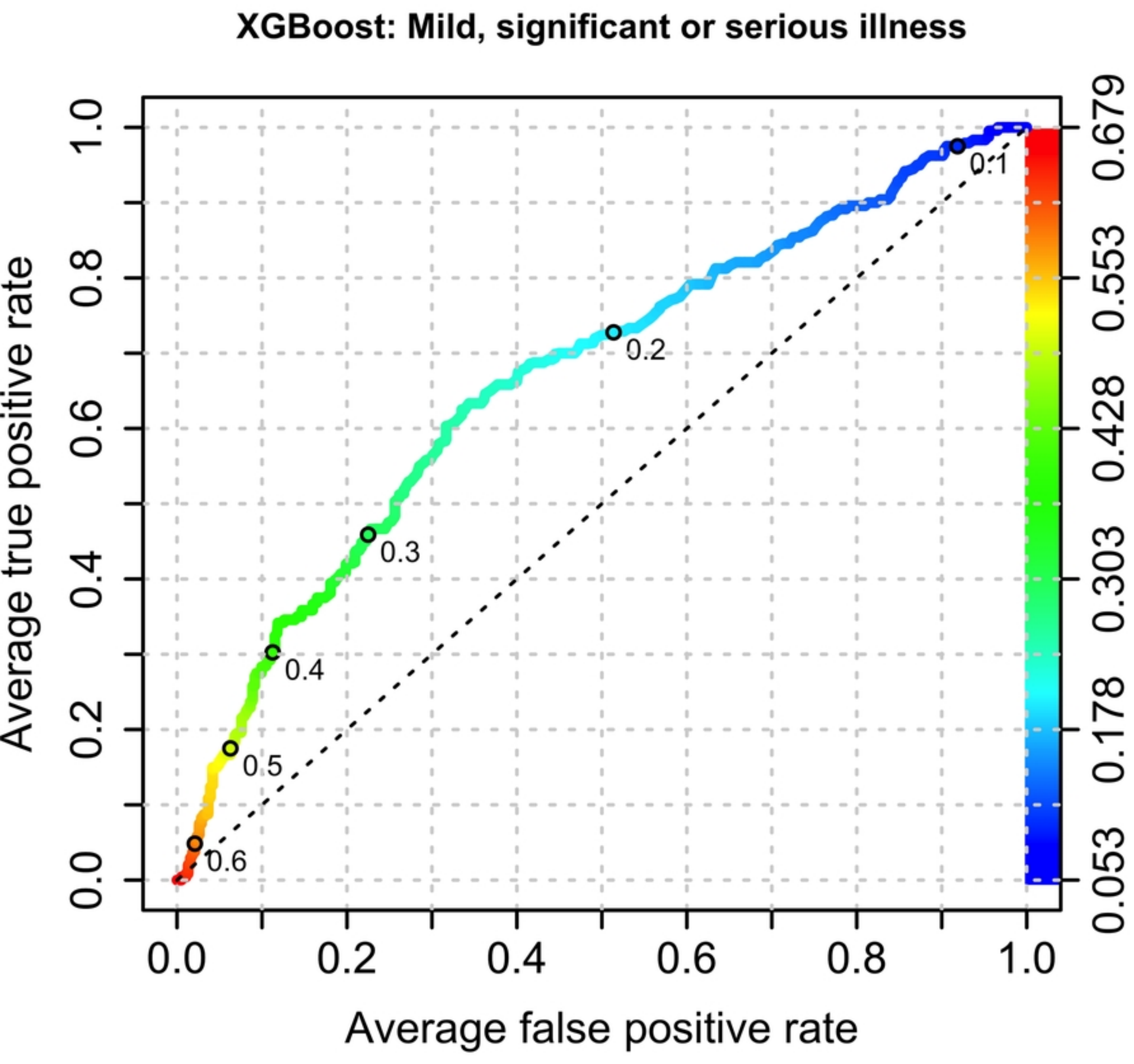
XGBoost: Mild, significant or serious illness
Average true positive rate
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.053
0.178
0.303
0.428
0.553
0.679

Figure 6a

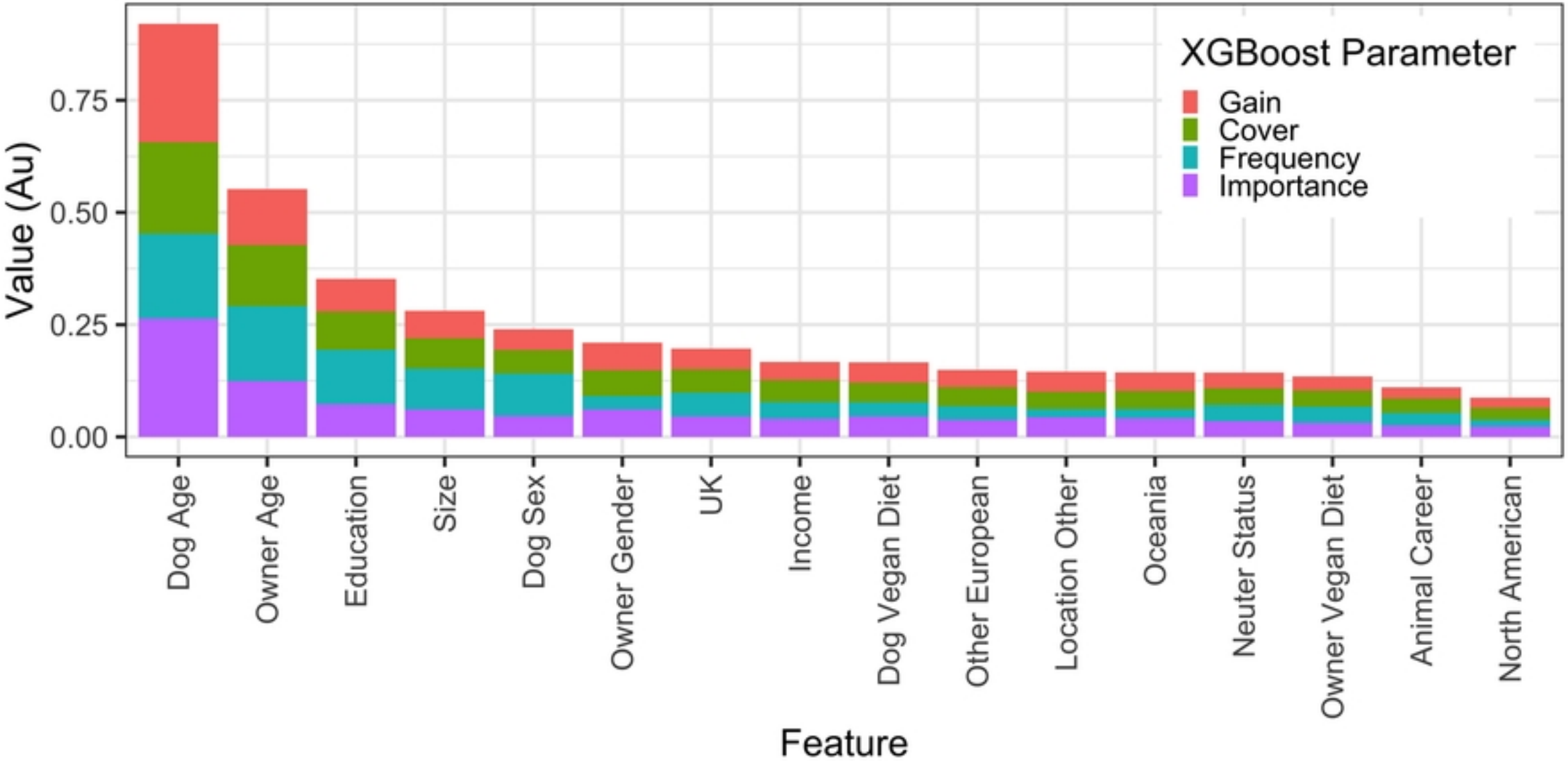
XGBoost Parameter
Gain
Cover
Frequency
Importance
Value (Au)
0.75
0.50
0.25
0.00
Dog Age
Owner Age
Education
Size
Dog Sex
Owner Gender
UK
Income
Dog Vegan Diet
Other European
Location Other
Oceania
Neuter Status
Owner Vegan Diet
Animal Career
North American
Feature

Figure 6b

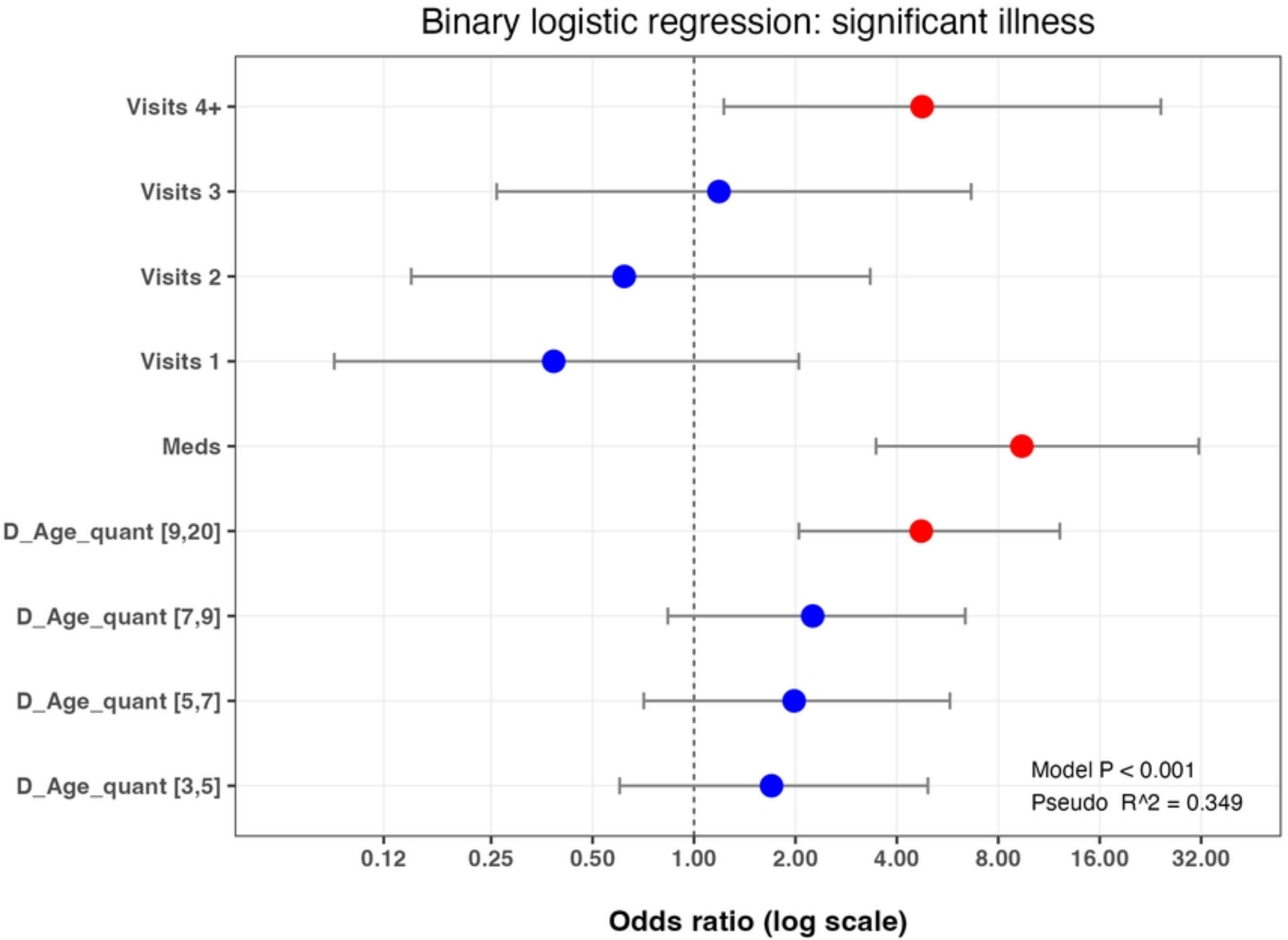
Binary logistic regression: significant illness
Visits 4+
Visits 3
Visits 2
Visits 1
Meds
D_Age_quant [9,20]
D_Age_quant [7,9]
D_Age_quant [5,7]
D_Age_quant [3,5]
Model P < 0.001
Pseudo R^2 = 0.349
0.12
0.25
0.50
1.00
2.00
4.00
8.00
16.00
32.00
Odds ratio (log scale)

Figure 7

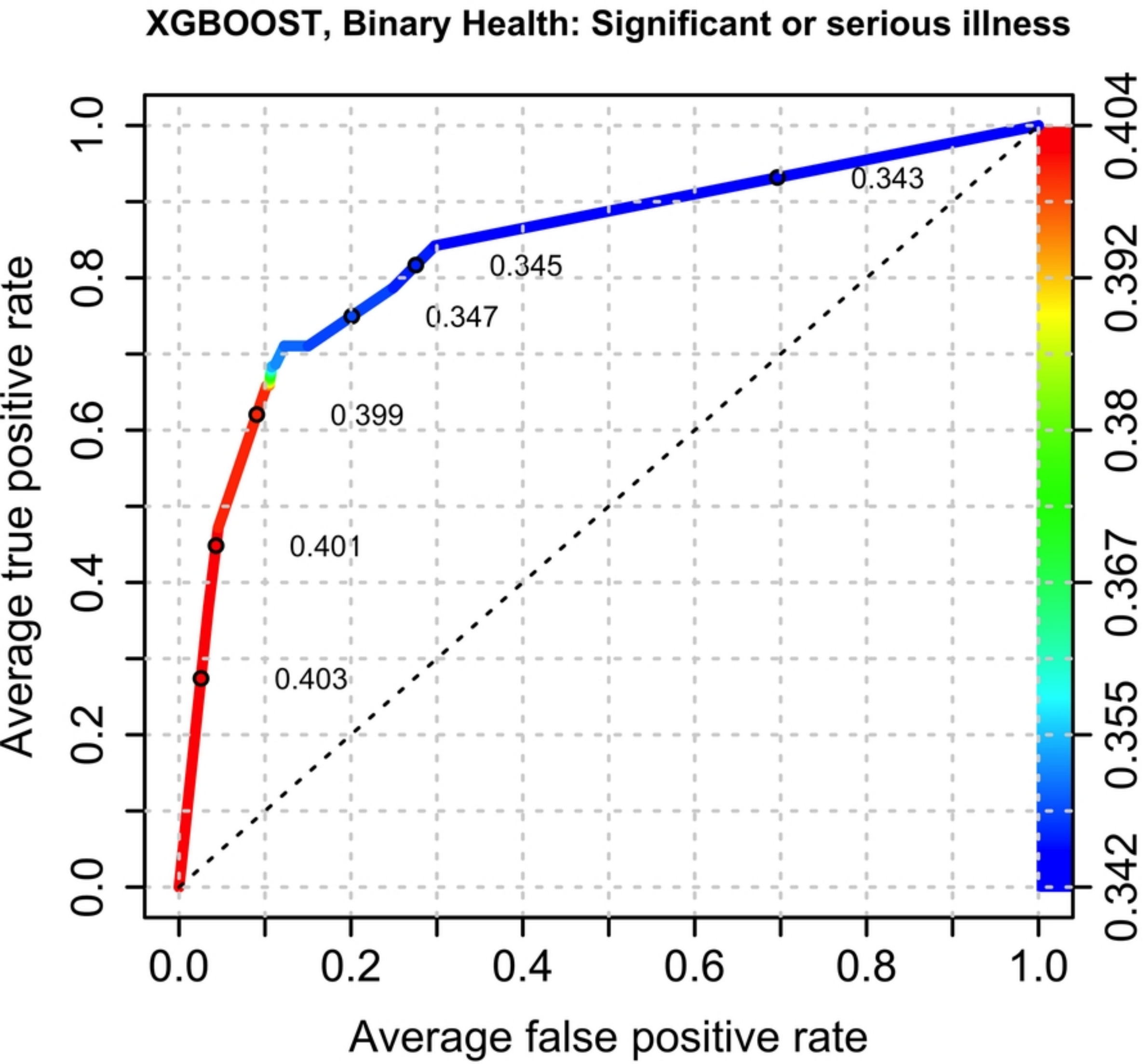
XGBOOST, Binary Health: Significant or serious illness
Average true positive rate
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.342
0.355
0.367
0.38
0.392
0.404
0.343
0.345
0.347
0.399
0.401
0.403

Figure 8a

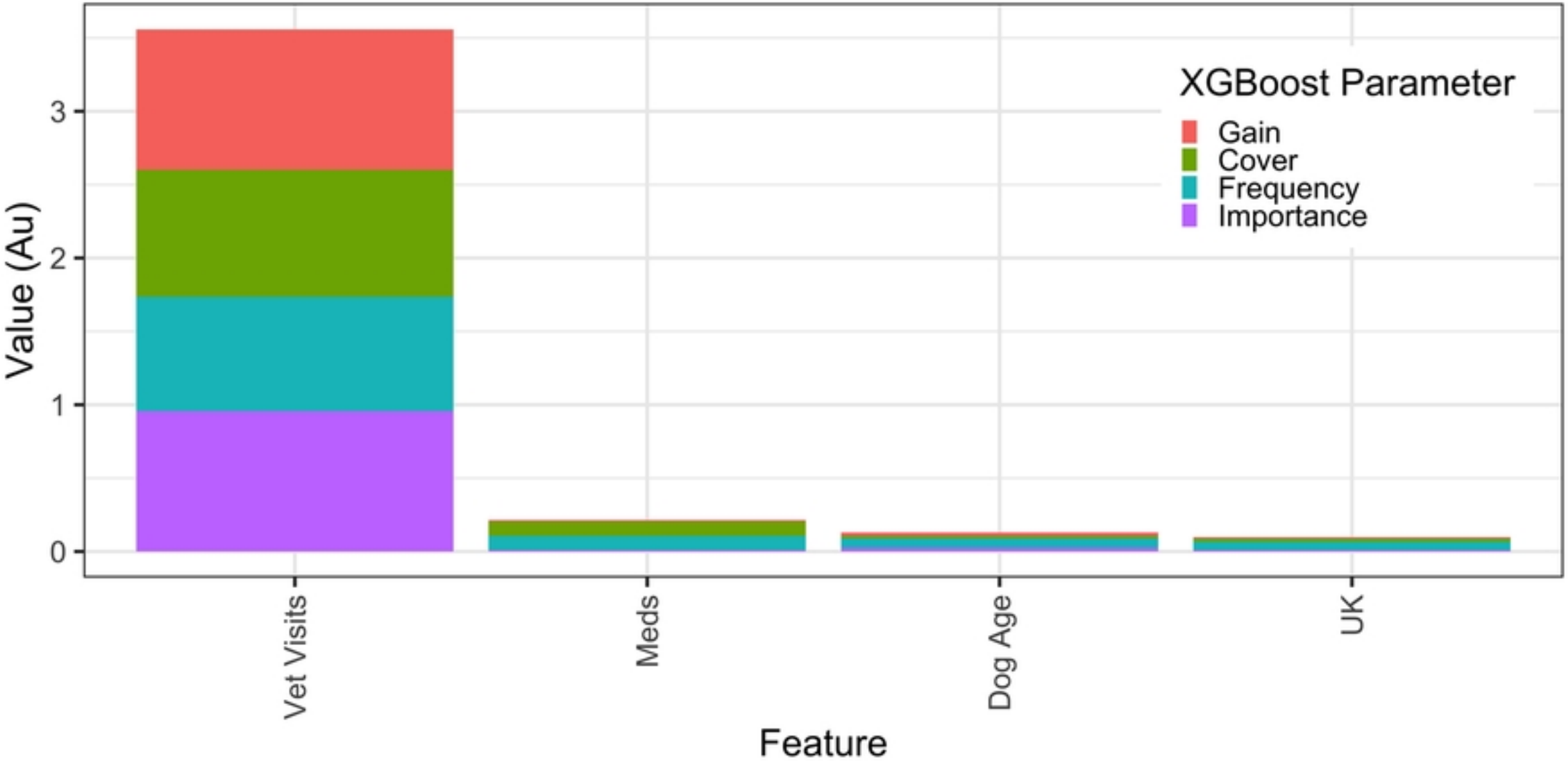
XGBoost Parameter
Gain
Cover
Frequency
Importance
Value (Au)
0
1
2
3
Vet Visits
Meds
Dog Age
UK
Feature

Figure 8b

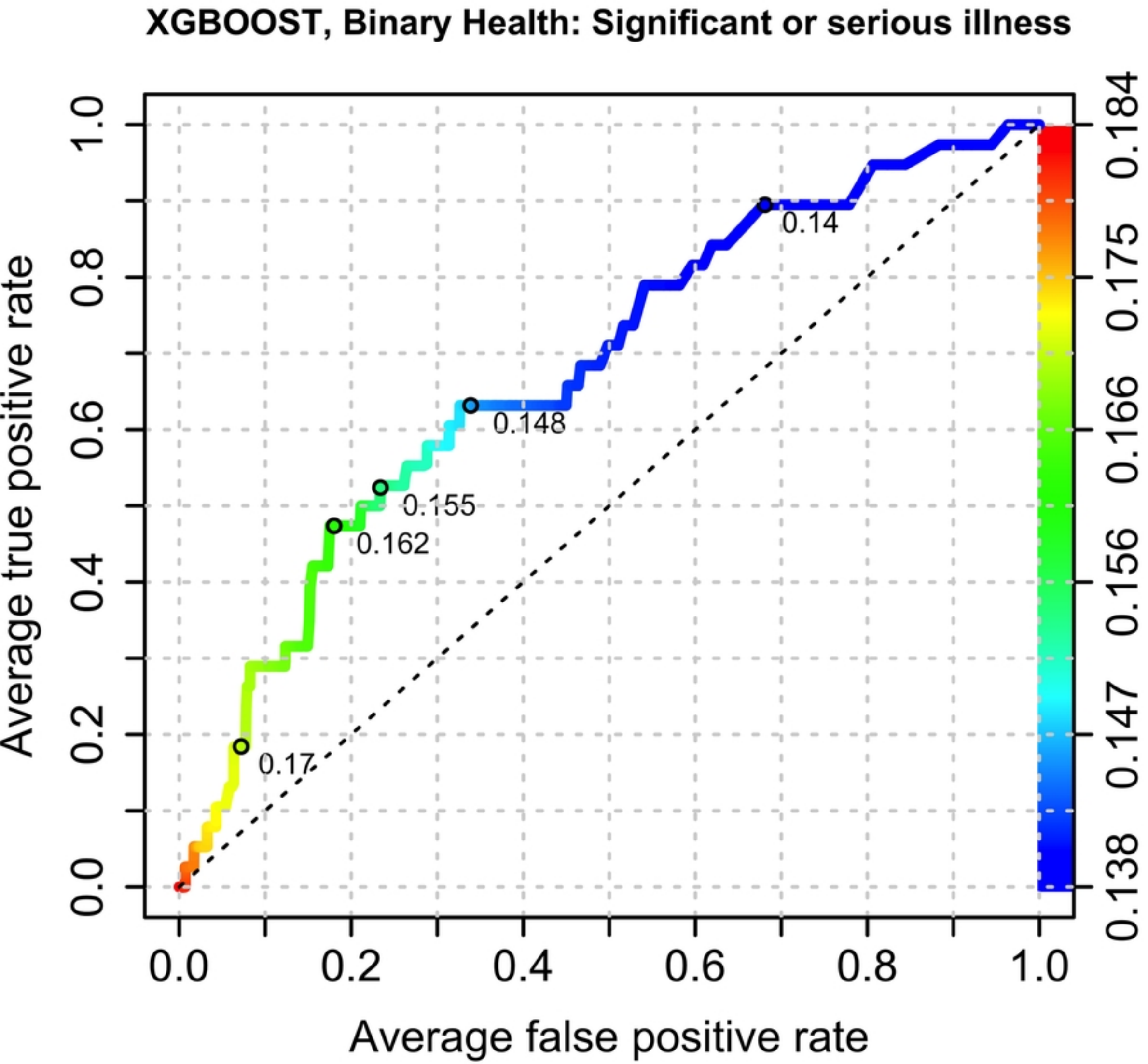
XGBOOST, Binary Health: Significant or serious illness
Average true positive rate
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.138
0.147
0.156
0.166
0.175
0.184
0.14
0.148
0.155
0.162
0.17

Figure 9a

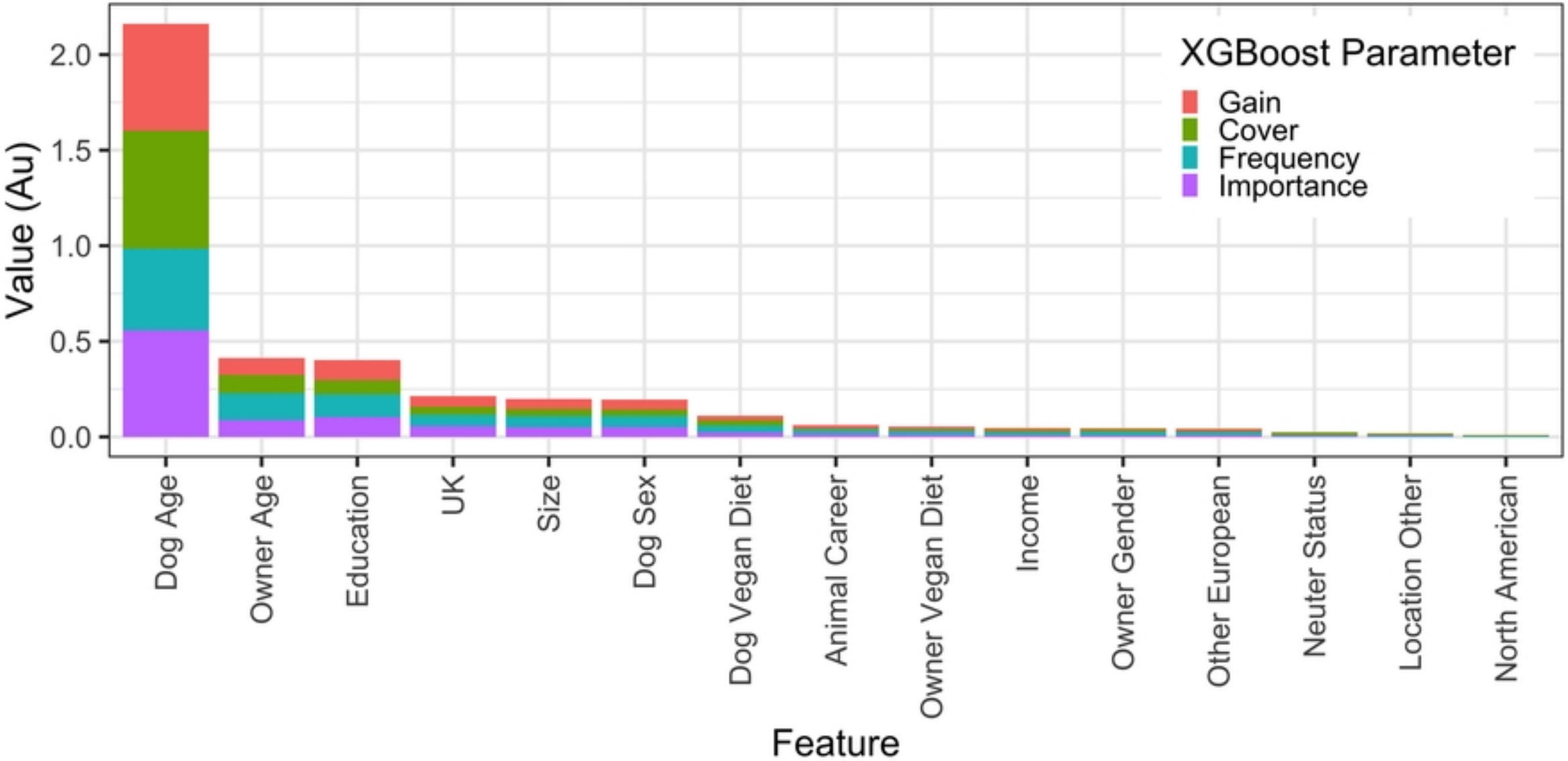
XGBoost Parameter
Gain
Cover
Frequency
Importance
Value (Au)
2.0
1.5
1.0
0.5
0.0
Dog Age
Owner Age
Education
UK
Size
Dog Sex
Dog Vegan Diet
Animal Career
Owner Vegan Diet
Income
Owner Gender
Other European
Neuter Status
Location Other
North American
Feature

Figure 9b