Weibel, C.J., M.R. Dasari, D.A. Jansen, L.R. Gesquiere, R.S. Mututua, J.K. Warutere, L.I. Siodi, S.C. Alberts, J. Tung, E.A. Archie. Using non-invasive behavioral and physiological data to measure biological age in wild baboons. *GeroScience.* 2024.

File list:

1. 1_data
    a. traits.csv
    b. backfill_data.csv
    c. rf_avg_predictions.csv
    d. enr_avg_predictions.csv
    e. gpm_avg_predictions.csv
    f. adversity_covariates.csv
2. 2_preparing_data
    a. 1_investigating_NAs.Rmd
    b. 2_prep_data_impute.Rmd
3. 3_machine_learning_models
    a. 1_random_forest_model.R
    b. 2_rf_predictions.Rmd
    c. 3_enr_model_predictions.Rmd
    d. 4_gpm_model_predictions.Rmd
    e. 5_model_comparisons.Rmd
4. 4_bio_implications
    a. 1_rf_aging_models.Rmd

File descriptions:

1a. 1_data/traits.csv

This file contains information on the 49 traits that exhibited significant linear or curvilinear associations with age, before filtering or data imputation. The data frame is set up such that each trait (and a few other relevant metrics) are shown in each column and each row represents a female's year of life. This data file will be read into the 2a script: "2_preparing_data/1_investigating_NAs.Rmd".

Variable name	Structure	Description
anon_sname	Factor	Anonymized names of subjects in the population
age_year	Integer	The female's age (in years) when the traits were measured
birth	Date	Subject's birth date
start_date	Date	Date that each female year of life began
end_date	Date	Date that each female year of life ended

stat_date	Date	The last date the female was observed (due to censoring or death)
avg_resid_log_gc	Numeric	Residuals of a linear model of log transformed **fecal glucocorticoid concentrations** controlling for variables shown in Table S2, averaged within a given year of a female's life
avg_resid_log_e	Numeric	Residuals of a linear model of log transformed **fecal estrogen concentrations** controlling for variables shown in Table S2, averaged within a given year of a female's life
avg_resid_log_p	Numeric	Residuals of a linear model of log transformed **fecal progesterone concentrations** controlling for variables shown in Table S2, averaged within a given year of a female's life
avg_resid_log_th	Numeric	Residuals of a linear model of log transformed **fecal thyroid hormone (T3) concentrations** controlling for variables shown in Table S2, averaged within a given year of a female's life
avg_log_trichuris_resid	Numeric	Residuals of a linear model of log transformed **trichuris count** controlling for variables shown in Table S2, averaged within a given year of a female's life
avg_richness_resid	Numeric	Residuals of a linear model of fecal **parasite richness** controlling for variables shown in Table S2, averaged within a given year of a female's life
avg_log_cycling_resid	Numeric	Residuals of a linear model of the subject's log transformed **cycling duration** controlling for variables shown in Table S2, averaged within a given year of a female's life if she began cycling more than once in that year
avg_preg_resid	Numeric	Residuals of a linear model of the subject's **pregnancy duration** controlling for variables shown in Table S2, averaged within a given year of a female's life if she had multiple pregnancies that were conceived in that year
any_misc	Factor	3-level factor depicting the subject's **pregnancy outcome** in a given year of life: • no_preg = the subject did not conceive • miscarriage = the subject conceived and experienced fetal loss • successful_preg = the subject conceived and gave birth to a live offspring

any_off_survive	Factor	3-level factor depicting the status of the subject's **offspring survival** in a given year of life: • no_off_birth = the subject did not give birth to a live offspring • off_died = the subject gave birth to a live offspring and it died before reaching 70 weeks of age • off_survived = the subject gave birth to a live offspring and it survived to at least 70 weeks
sci_f_dir_resid	Numeric	Residuals of a linear model of female social connectedness to other females based on grooming behaviors that the female initiated (**SCI directed (F)**) in a given year of life controlling for variables shown in Table S2
sci_f_rec_resid	Numeric	Residuals of a linear model of female social connectedness to other females based on grooming behaviors that the female received (**SCI received (F)**) in a given year of life controlling for variables shown in Table S2
sci_f_resid	Numeric	Residuals of a linear model of female social connectedness to other females based on grooming behaviors that the female initiated and received (**SCI (F)**) in a given year of life controlling for variables shown in Table S2
sci_m_dir_resid	Numeric	Residuals of a linear model of female social connectedness to males based on grooming behaviors that the female initiated (**SCI directed (M)**) in a given year of life controlling for variables shown in Table S2
sci_m_rec_resid	Numeric	Residuals of a linear model of female social connectedness to males based on grooming behaviors that the female received (**SCI received (M)**) in a given year of life controlling for variables shown in Table S2
agi_f_dir_resid	Numeric	Residuals of a linear model of the relative amount of agonisms the subject directed at other females (**AGI directed (F)**) in a given year of life controlling for variables shown in Table S2
agi_f_rec_resid	Numeric	Residuals of a linear model of the relative amount of agonisms the subject received from other females (**AGI received (F)**) in a given year of life controlling for variables shown in Table S2
agi_f_resid	Numeric	Residuals of a linear model of the relative amount of agonisms the subject received from and directed at other females (**AGI (F)**)

		in a given year of life controlling for variables shown in Table S2
agi_m_rec_resid	Numeric	Residuals of a linear model of the relative amount of agonisms the subject received from males (**AGI received (M)**) in a given year of life controlling for variables shown in Table S2
agi_m_resid	Numeric	Residuals of a linear model of the relative amount of agonisms the subject received from and directed at males (**AGI (M)**) in a given year of life controlling for variables shown in Table S2
dsi_f_resid	Numeric	Residuals of a linear model of the relative strength of the subject's bonds with their top three female grooming partners in a given year of life (**DSI (F)**) controlling for variables shown in Table S2
recip_f_resid	Numeric	Residuals of a linear model of a measure of how reciprocal the subject's grooming relationships are with females in a given year of life (**reciprocity (F)**) controlling for variables shown in Table S2
recip_m_resid	Numeric	Residuals of a linear model of a measure of how reciprocal the subject's grooming relationships are with males in a given year of life (**reciprocity (M)**) controlling for variables shown in Table S2
not_bonded_m_resid	Numeric	Residuals of a linear model of the number of males in the subject's social group that the subject did not have bonds with in a given year of life (**no. not bonded (M)**) controlling for variables shown in Table S2
not_bonded_f_resid	Numeric	Residuals of a linear model of the number of other females in the subject's social group that the subject did not have bonds with in a given year of life (**no. not bonded (F)**) controlling for variables shown in Table S2
strongly_bonded_f_resid	Numeric	Residuals of a linear model of the number of other females in the subject's social group that the subject had a strong bond with in a given year of life (**no. strong bonds (F)**) controlling for variables shown in Table S2
weakly_bonded_f_resid	Numeric	Residuals of a linear model of the number of other females in the subject's social group that the subject had a weak bond with in a given year of life (**no. weak bonded (F)**) controlling for variables shown in Table S2
total_grooming_partners_f_resid	Numeric	Residuals of a linear model of the total number of female grooming partners the subject had in a given year of life (**no. total**

		grooming partners (F)) controlling for variables shown in Table S2
total_grooming_partners_resid	Numeric	Residuals of a linear model of the total number of grooming partners the subject had in a given year of life (**no. total grooming partners (M+F)**) controlling for variables shown in Table S2
eigen_resid	Numeric	Residuals of a linear model of a measure of influence the subject had on their social network in a given year of life (**eigenvector centrality**) controlling for variables shown in Table S2
avg_mom_rel_rank	Factor	3-level factor depicting the subject's **rank relative to her mother** in a given year of life: • at_moms_rank = the subject ranks the same as her mother • below_moms_rank = the subject ranks below her mother • above_moms_rank = the subject ranks above her mother
rank_rel_any_daughter	Factor	3-level factor depicting the subject's **rank relative to her daughter** in a given year of life: • no_alive_daughter = the subject does not have any living daughters • focal_above_daughters_rank = the subject ranks above all her daughters • focal_below_daughters_rank = the subject ranks below at least one of her daughters
cat_abs_active_rank_change	Factor	3-level factor depicting the subject's **active rank change** in a given year of life: • no_change = subject did not experience any active rank reversals • active_rise = subject experienced an active rise in rank • active_fall = subject experienced an active fall in rank
per_groom_resid	Numeric	Residuals of a linear model of the percent of time the subject spent grooming others in a given year of life when she had a young infant present (**per. time grooming (IP)**) controlling for variables shown in Table S2
per_be_groomed_infant_resid	Numeric	Residuals of a linear model of the percent of time the subject spent being groomed by her infant in a given year of life (**per. time being groomed by infant**) controlling for variables shown in Table S2

per_kc_dorsal_resid	Numeric	Residuals of a linear model of the percent of time the subject spent supporting her infant dorsally in a given year of life (**per. time supporting infant dorsally**) controlling for variables shown in Table S2
per_kc_ventral_resid	Numeric	Residuals of a linear model of the percent of time the subject spent supporting her infant ventrally in a given year of life (**per. time supporting infant dorsally**) controlling for variables shown in Table S2
per_kc_support_resid	Numeric	Residuals of a linear model of the percent of time the subject spent supporting her infant in a given year of life (**per. time supporting infant**) controlling for variables shown in Table S2
per_kc_away_none_resid	Numeric	Residuals of a linear model of the percent of time the subject spent >3 meter away from her infant in a given year of life (**per. time away from infant**) controlling for variables shown in Table S2
per_ks_suckle_resid	Numeric	Residuals of a linear model of the percent of time the infant spent sucking from the subject in a given year of life (**per. time infant suckling**) controlling for variables shown in Table S2
per_no_neighbors_resid	Numeric	Residuals of a linear model of the percent of time the subject spent with no close neighbors in a given year of life when she had a young infant present (**per. time no neighbors (IP)**) controlling for variables shown in Table S2
ni_per_groom_resid	Numeric	Residuals of a linear model of the percent of time the subject spent grooming others in a given year of life when she did not have a young infant present (**per. time grooming (NIP)**) controlling for variables shown in Table S2
ni_per_other_soc_resid	Numeric	Residuals of a linear model of the percent of time the subject spent engaging in other social behaviors in a given year of life when she did not have a young infant present (**per. time other social behaviors (NIP)**) controlling for variables shown in Table S2
ni_per_rest_resid	Numeric	Residuals of a linear model of the percent of time the subject spent resting in a given year of life when she did not have a young infant present (**per. time resting (NIP)**) controlling for variables shown in Table S2
ni_per_forage_resid	Numeric	Residuals of a linear model of the percent of time the subject spent foraging in a given

		year of life when she did not have a young infant present (**per. time foraging (NIP)**) controlling for variables shown in Table S2
ni_per_no_neighbors_resid	Numeric	Residuals of a linear model of the percent of time the subject spent with no close neighbors in a given year of life when she did not have a young infant present (**per. time no neighbors (NIP)**) controlling for variables shown in Table S2
major_health_event_cat	Factor	Binary variable representing whether the subject experienced a major health event in a given year of life (1 = yes, 0 = no)
limp_cat	Factor	Binary variable representing whether the subject experienced a limp in a given year of life (1 = yes, 0 = no)
puncture_cat	Factor	Binary variable representing whether the subject experienced a puncture in a given year of life (1 = yes, 0 = no)

1b. 1_data/backfill_data.csv

This file contains information on the 49 traits that exhibited significant linear or curvilinear associations with age, after filtering and forward-filling and back-filling to provide values for missing female-trait-years. This dataset is generated by the 2a script: "2_preparing_data/1_investigating_NAs.Rmd" and is the input for the 2b script: "2_preparing_data/2_prep_data_impute.Rmd". All column descriptions can be found in the 1a description above.

1c. 1_data/rf_avg_predictions.csv

This file contains the predictions from the random forest model for each of the five imputation data sets, along with the average predictions across all five imputations. This dataset is generated by the 3a and 3b scripts, "3_machine_learning_models/1_random_forest_model.R" and "3_machine_learning_models/2_rf_predictions.Rmd", respectively.

Variable name	**Structure**	**Description**
anon_sname	Factor	Anonymized names of subjects in the population
age_year	Integer	The female's age (in years) when the traits were measured
bio_age1	Numeric	Age prediction from the random forest model for the 1st imputation data set
bio_age2	Numeric	Age prediction from the random forest model for the 2nd imputation data set
bio_age3	Numeric	Age prediction from the random forest model for the 3rd imputation data set
bio_age4	Numeric	Age prediction from the random forest model for the 4th imputation data set

bio_age5	Numeric	Age prediction from the random forest model for the 5th imputation data set
avg_bio_age	Numeric	Average age prediction from the random forest model across the five imputation data sets

1d. 1_data/enr_avg_predictions.csv

This file contains the predictions from the elastic net regression for each of the five imputation data sets, along with the average predictions across all five imputations. This dataset is generated by the 3c script: "3_machine_learning_models/3_enr_model_predictions.Rmd". The dataset contains the same columns as the 1c table above, save for the model type.

1e. 1_data/gpm_avg_predictions.csv

This file contains the predictions from the gaussian process model for each of the five imputation data sets, along with the average predictions across all five imputations. This dataset is generated by the 3d script:
"3_machine_learning_models/4_gpm_model_predictions.Rmd". The dataset contains the same columns as the 1c table above, save for the model type.

1f. 1_data/adversity_covariates.csv

This file contains early adversity data and other covariates used as predictor variables in the survival models and linear models that assess the relationship between early adversity and biological age. This data set is needed to run the 4a script:
"4_bio_implications/1_rf_aging_models.Rmd".

Variable name	Structure	Description
anon_sname	Factor	Anonymized names of subjects in the population
age_year	Integer	The female's age (in years) when the traits were measured
birth	Date	Subject's birth date
start	Date	Date that each female year of life began
end	Date	Date that each female year of life ended
statdate	Date	The last date the female was observed (due to censoring or death)
status	Factor	The status of the subject on their statdate
age_at_statdate	Numeric	The subject's age in years when she died or was censored
mom_rank_birth	Integer	The subject's mother's ordinal rank at birth
group_size_birth	Integer	The subject's group size at birth
mom_sci	Numeric	The subject's mother's level of social isolation in the first two years of life
mom_death_cat	Factor	Whether the subject's mother died (0 = no death; 1 = death)
competing_sibling_cat	Factor	Whether the subject had a competing

		younger sibling (0 = no sibling; 1 = sibling)
drought_cat	Factor	Whether the subject experienced a drought in the first year of life (0 = no drought; 1 = drought)
mom_rank_birth_cat	Factor	Whether the subject's mother was in the bottom quartile of female ranks (ordinal rank >= 12)
group_size_birth_cat	Factor	Whether the subject's group size was in the top quartile of group size (>=36)
mom_sci_cat	Factor	Whether the subject's mother was in the bottom quartile of social connectedness (<=0.325)
cumulative_adversity	Integer	The number of adverse events the subject experienced in early life (range = 0-6)
cumulative_adversity_three_plus	Integer	The number of adverse events the subject experienced in early life. Subjects with three or more adverse events are grouped together (range = 0-3)
avg_ordrank	Numeric	The subject's average ordinal rank across each month in a given year of life
avg_proprank	Numeric	The subject's average proportional rank across each month in a given year of life
rain_anom	Numeric	How relatively wet or dry the conditions were in a given year of life
avg_group_size	Numeric	The subject's average group size across a given year of life

2a. 2_preparing_data/1_investigating_NAs.Rmd

This script reads in the 1a data frame ("1_data/traits.csv"). It calculates the number of missing traits for each female's year of life and then fills in missing values with data from the year immediate before or after for each female. The code then filters for all female years of life when 34 or fewer traits were missing.

2b. 2_preparing_data/2_prep_data_impute.Rmd

This script reads in the 1b data frame ("1_data/backfill_data.csv"). It splits the data into 5 training and 5 test data sets, imputes missing data for each data set 5 times, and then standardizes each of the traits.

3a. 3_machine_learning_models/1_random_forest_models.R

This script uses the training and test data frames generated in the 2b script ("2_preparing_data/2_prep_data_impute.Rmd") and uses a random forest model to predict biological age for each of the data sets.

3b. 3_machine_learning_models/2_rf_predictions.Rmd

This script uses the model output and predictions from the 3a script ("3_machine_learning_models/1_random_forest_models.R"). It combines the predictions to calculate the average age prediction for each female year of life and also calculates the importance of each trait in the model.

3c. 3_machine_learning_models/3_enr_model_predictions.Rmd

This script uses the training and test data frames generated in the 2b script ("2_preparing_data/2_prep_data_impute.Rmd") and uses an elastic net regression to predict biological age for each of the data sets. It also combines the predictions to calculate the average age prediction for each female year of life.

3d. 3_machine_learning_models/4_gpm_model_predictions.Rmd

This script uses the training and test data frames generated in the 2b script ("2_preparing_data/2_prep_data_impute.Rmd") and uses a gaussian process model to predict biological age for each of the data sets. It also combines the predictions to calculate the average age prediction for each female year of life.

3e. 3_machine_learning_models/5_model_comparisons.Rmd

This script reads in data frames 1c, 1d, and 1e (outputs from the 3b, 3c, and 3d scripts). It calculates repeatability for each modeling technique and compares the performance of the three models.

4a. 4_bio_implications/1_rf_aging_models.Rmd

This script reads in the 1c data frame (the output from the 3b script; "1_data/rf_avg_predictions.csv"), as well as the 1f data frame ("1_data/adversity_covariates.csv"). It evaluates the relationship between early-life adversity and biological age, as well as the relationship between biological age, early-life adversity, and survival.

Additional note:

Data and code for testing which of the original 78 traits change with age and visualizing the relationship that the 49 significant traits had with age is not currently included in this repository, but can be provided upon request.