Supplementary Material

Supplementary dataset information

Supplementary Table S1. Imaging characteristics of institutional datasets (1/2/3) used in this study.

Variables	Total (n = 8926)	Dataset1 (n = 4274)	Dataset2 (n = 1996)	Dataset3 (n = 2656)	Statistic	*P*
Manufacturer, n (%)					χ^2=5000.25	<0.001
GE MEDICAL SYSTEMS	4522 (50.66)	3592 (84.04)	459 (23.00)	471 (17.73)		
NMS	858 (9.61)	541 (12.66)	36 (1.80)	281 (10.58)		
Philips	1484 (16.63)	5 (0.12)	481 (24.10)	998 (37.58)		
SIEMENS	2062 (23.10)	136 (3.18)	1020 (51.10)	906 (34.11)		
Scanner model, n (%)					χ^2=7134.53	<0.001
Discovery CT750 HD	2455 (27.50)	2308 (54.00)	105 (5.26)	42 (1.58)		
iCT 256	1484 (16.63)	5 (0.12)	481 (24.10)	998 (37.58)		
NeuViz Prime	858 (9.61)	541 (12.66)	36 (1.80)	281 (10.58)		
Optima CT680 Expert	512 (5.74)	80 (1.87)	92 (4.61)	340 (12.80)		
Revolution CT	1555 (17.42)	1204 (28.17)	262 (13.13)	89 (3.35)		
SOMATOM Definition Flash	1451 (16.26)	0 (0.00)	552 (27.66)	899 (33.85)		
SOMATOM Drive	7 (0.08)	7 (0.16)	0 (0.00)	0 (0.00)		

Variables	Total (n = 8926)	Dataset1 (n = 4274)	Dataset2 (n = 1996)	Dataset3 (n = 2656)	Statistic	*P*
SOMATOM Force	604 (6.77)	129 (3.02)	468 (23.45)	7 (0.26)		
Phase, n(%)					χ^2=79.06	<0.001
AP	1920 (21.51)	825 (19.30)	414 (20.74)	681 (25.64)		
DP	2609 (29.23)	1309 (30.63)	519 (26.00)	781 (29.41)		
NOC	2184 (24.47)	1013 (23.70)	509 (25.50)	662 (24.92)		
PVP	2213 (24.79)	1127 (26.37)	554 (27.76)	532 (20.03)		
Slice Thickness , n(%)					χ^2=5276.91	<0.001
1	2482 (27.81)	632 (14.79)	767 (38.43)	1083 (40.78)		
1.25	4093 (45.85)	3589 (83.97)	268 (13.43)	236 (8.89)		
3	53 (0.59)	53 (1.24)	0 (0.00)	0 (0.00)		
5	2298 (25.75)	0 (0.00)	961 (48.15)	1337 (50.34)		

DSC, Dice similarity coefficient; VS, volume similarity; HD, Hausdorff distance; Q1/Q3, first/third quartile.

Supplementary Table S2. Imaging phase and slice thickness distributions of public CT datasets used for external validation

Variables	Total (n = 2809)	Abdomen_CT_date_has_label (n = 944)	Abdomen_CT-1k_Fullysupervised_subtask1 (n = 321)	Abdomen_CT-1k_Semi (n = 38)	AMOS (n = 270)	Flare23 (n = 1121)	RSNA (n = 115)
Phase, n (%)							
AP	860 (30.62)	285 (30.19)	87 (27.19)	4 (10.53)	73 (27.04)	371 (33.10)	40 (34.78)
DP	379 (13.49	148 (15.68)	54 (16.88)	16 (42.11)	21 (7.78)	127 (11.3	13 (11.3

Variables	Total (n = 2809)	Abdomen_CT_date_has_label (n = 944)	Abdomen_CT-1k_Fullysupervised_subtask1 (n = 321)	Abdomen_CT-1k_Semi (n = 38)	AMOS (n = 270)	Flare23 (n = 1121)	RSNA (n = 115)
	)					3)	0)
NOC	146 (5.20)	7 (0.74)	0 (0.00)	3 (7.89)	128 (47.41)	7 (0.62)	1 (0.87)
PVP	1424 (50.69)	504 (53.39)	180 (55.94)	15 (39.47)	48 (17.78)	616 (54.95)	61 (53.04)
Slice thickness (mm), n (%)							
<1.00	250 (8.90)	106 (11.23)	2 (0.62)	0 (0.00)	0 (0.00)	114 (10.17)	28 (24.35)
>5.00	14 (0.50)	5 (0.53)	1 (0.31)	3 (7.89)	0 (0.00)	5 (0.45)	0 (0.00)
1.00	387 (13.78)	132 (13.98)	72 (22.50)	0 (0.00)	0 (0.00)	180 (16.06)	3 (2.61)
1.25	52 (1.85)	10 (1.06)	3 (0.94)	0 (0.00)	5 (1.85)	22 (1.96)	12 (10.43)
1.50	78 (2.78)	25 (2.65)	2 (0.62)	5 (13.16)	0 (0.00)	31 (2.77)	15 (13.04)
2.00	107 (3.81)	19 (2.01)	1 (0.31)	0 (0.00)	60 (22.22)	18 (1.61)	9 (7.83)
2.50	1005 (35.78)	357 (37.82)	193 (60.31)	3 (7.89)	7 (2.59)	425 (37.91)	19 (16.52)
3.00	168 (5.98)	61 (6.46)	6 (1.88)	1 (2.63)	5 (1.85)	70 (6.24)	25 (21.74)
3.75	14 (0.50)	6 (0.64)	2 (0.62)	0 (0.00)	0 (0.00)	6 (0.54)	0 (0.00)
4.00	34 (1.21)	14 (1.48)	2 (0.62)	4 (10.53)	0 (0.00)	14 (1.25)	0 (0.00)

Variables	Total (n = 2809)	Abdomen_CT_date_has_label (n = 944)	Abdomen_CT-1k_Fullysupervised_subtask1 (n = 321)	Abdomen_CT-1k_Semi (n = 38)	AMOS (n = 270)	Flare23 (n = 1121)	RSNA (n = 115)
5.00	700 (24.92)	209 (22.14)	36 (11.25)	22 (57.89)	193 (71.48)	236 (21.05)	4 (3.48)

NOC, non-contrast; AP, arterial phase; PVP, portal venous phase; DP, delayed phase; CT, computed tomography.

Supplementary Table S3. Spleen segmentation performance across public CT datasets in external validation

Variables	Total (n = 2809)	Abdomen_CT_date_has_label (n = 944)	Abdomen_CT-1k_Fullysupervised_subtask1 (n = 321)	Abdomen_CT-1k_Semi (n = 38)	AMOS (n = 270)	Flare23 (n = 1121)	RSNA (n = 115)
DSC, median (Q1, Q3)	0.98 (0.97, 0.99)	0.98 (0.98, 0.99)	0.98 (0.98, 0.99)	0.98 (0.98, 0.99)	0.97 (0.97, 0.98)	0.98 (0.97, 0.99)	0.98 (0.97, 0.98)
VS, median (Q1, Q3)	0.99 (0.99, 1.00)	0.99 (0.99, 1.00)	1.00 (0.99, 1.00)	0.99 (0.99, 1.00)	0.99 (0.99, 1.00)	0.99 (0.99, 1.00)	0.99 (0.98, 1.00)
HD (mm), median (Q1, Q3)	0.02 (0.01, 0.03)	0.02 (0.01, 0.02)	0.02 (0.01, 0.02)	0.02 (0.01, 0.02)	0.03 (0.02, 0.04)	0.02 (0.01, 0.03)	0.02 (0.02, 0.04)

DSC, Dice similarity coefficient; VS, volume similarity; HD, Hausdorff distance。

Supplementary methods and results

1. Prediction model development and evaluation

Linear regression models were constructed using 578 cases to identify predictors of spleen volume from a set of candidate anthropometric and demographic variables, including body surface area (BSA), weight, height, body mass index (BMI), and age.

1.1 Descriptive statistics and data distribution

Demographic and anthropometric characteristics of the study cohort (n = 578) were summarized. The dataset contained no missing data. Continuous variables—age, spleen volume, body height, body weight, body mass index (BMI), and body surface area (BSA)—were described using standard descriptive statistics: sample size, mean, standard deviation (SD), median, interquartile range (IQR), and range.

The normality of spleen volume distribution was assessed using the Shapiro-Wilk test, supplemented by calculations of skewness and kurtosis.

A correlation matrix was constructed to examine pairwise relationships among candidate predictor variables (age, body height, body weight, BMI, BSA) using Pearson correlation coefficients. Correlations were calculated from complete observations to assess potential multicollinearity concerns for subsequent regression modeling.

1.2 Model development and variable selection

Predictor selection for modeling spleen volume was performed using bidirectional stepwise regression with Akaike Information Criterion (AIC) optimization. Starting from a null model (intercept-only), the algorithm sequentially added or removed variables from the candidates based on AIC improvement, continuing until no further AIC reduction was achievable. The search space was bounded by the null model (lower bound) and the full model containing all five predictors (upper bound).

Regression coefficients, standard errors, 95% confidence intervals, t-statistics, and p-values were extracted from the final spleen volume prediction model (Volume ~ BSA + Age). Confidence intervals were computed using the t-distribution with 575 degrees of freedom.

1.3 Model diagnostics and transformation

Overall model performance of the final linear regression model (Volume ~ BSA + Age) was evaluated using standard goodness-of-fit metrics: R-squared, adjusted R-squared, and the overall F-test. R-squared measures the proportion of variance in spleen volume explained by the predictors. Adjusted R-squared accounts for the number of predictors. The F-statistic tests the null hypothesis that all regression coefficients (excluding the intercept) equal zero.

Multicollinearity among candidate predictors (Age, Height, Weight, BMI, BSA) was assessed using variance inflation factors (VIFs). VIF values exceeding 5 indicate moderate multicollinearity, while values exceeding 10 indicate severe multicollinearity.

Three key linear regression assumptions were formally tested. 1) Normality of residuals was assessed using the Kolmogorov-Smirnov (KS) test, which compares the empirical distribution of residuals to a normal distribution. 2) Homoscedasticity (constant variance of residuals) was tested using the Breusch-Pagan test, which regresses squared residuals on the predictors. 3) Independence of residuals (no autocorrelation) was evaluated using the Durbin-Watson test, with a statistic near 2 indicating no autocorrelation. Significance was evaluated at $\alpha = 0.05$.

To address the observed violations of normality and homoscedasticity, a natural log-transformation was applied to the response variable (log Volume = log(Volume)). The same linear model structure was then fitted (log Volume ~ BSA + Age). Model coefficients were

exponentiated to provide interpretation on the original spleen volume scale. Specifically, exp(β) represents the multiplicative change in the expected spleen volume associated with a one-unit increase in the corresponding predictor, holding other variables constant. Both point estimates and 95% confidence intervals were back-transformed via exponentiation, yielding interpretable multiplicative effects on the original volume scale.

1.4 Internal validation and calibration

A bootstrap procedure with 1,000 replicates was performed to internally validate the log-transformed linear regression model (Log_Volume ~ BSA + Age) and correct for optimism. In each bootstrap sample, the model was refitted, and performance was evaluated both in-sample (apparent) and out-of-bag (OOB). The optimism—defined as the average difference between apparent and OOB performance—was computed for R^2 and root mean square error (RMSE). Optimism-corrected metrics were obtained by subtracting the mean optimism from apparent performance. 95% confidence intervals for the corrected metrics were derived from the bootstrap distribution.

Model calibration was assessed to evaluate the agreement between predicted and observed spleen volumes on the log-transformed scale. Calibration was examined graphically by plotting observed versus predicted log-transformed spleen volumes and fitting a linear regression of observed values on predicted values. The calibration intercept reflects systematic bias (over- or under-prediction), while the calibration slope reflects potential over- or under-dispersion of predictions. Because calibration assessed on the same data used for model development is inherently optimistic, this analysis was considered descriptive rather than confirmatory. Calibration metrics were therefore interpreted cautiously and in conjunction with bootstrap-based internal validation results.

1.5 Sample size justification and model stability assessment

The model's stability was assessed through two key metrics: (1) the events-per-variable(EPV) ratio (n/p), calculated as the effective sample size (n = 578) divided by the number of predictors including intercept (p = 3), and (2) a bootstrap-derived shrinkage factor(1000 replicates). The shrinkage factor estimates the degree of overfitting by comparing the apparent model performance to its performance when applied to bootstrap samples. A shrinkage factor ≥0.9 indicates minimal overfitting.

1.6 Sensitivity analyses

To assess the necessity of including both BSA and Age in the final model, three competing models were compared: (1) BSA-only model, (2) Age-only model, and (3) the original BSA + Age model. Model performance was evaluated using Akaike Information Criterion (AIC), R-squared, mean absolute error (MAE), and root mean squared error (RMSE). Residual distributions were examined to assess calibration.

1.7. Subgroup analyses

A multiplicative interaction model was fitted to evaluate whether the effects of BSA and Age on log-transformed spleen volume differed by sex (Log_Volume ~ BSA * Sex + Age * Sex). Model comparisons were made via: 1) ANOVA F-test comparing the interaction model to the additive model (BSA + Age only); 2) Drop-in-deviance F-tests for each interaction term; and 3)

Examination of interaction coefficient estimates with t-tests.

To further validate the absence of sex-specific effects, three models were compared: 1) A pooled model (Log_Volume ~ BSA + Age) assuming identical effects across sexes; 2) A stratified model allowing all parameters (intercept and slopes) to differ by sex (Log_Volume ~ Sex/(BSA + Age) - 1); and 3) Sex-specific models fit separately to male and female subgroups. An F-test was used to compare the pooled versus stratified models, testing the null hypothesis that regression coefficients (intercept, BSA slope, Age slope) are equal between sexes.

Results

Prediction model development and evaluation

1. Participant characteristics and data distribution

1.1 Descriptive statistics

The cohort (n = 578) had a mean age of 54.1 ± 13.7 years, with a median spleen volume of 178 mL (IQR: 93.7). Mean height was 167 ± 7.8 cm, mean weight 68.0 ± 11.9 kg, mean BMI 24.4 ± 3.34 kg/m², and mean BSA 1.85 ± 0.181 m². Distributions were largely symmetric, as means and medians were comparable across variables. Spleen volume showed substantial variability (range: 51.1–644 mL), consistent with biological heterogeneity in organ size (Supplementary Table S4).

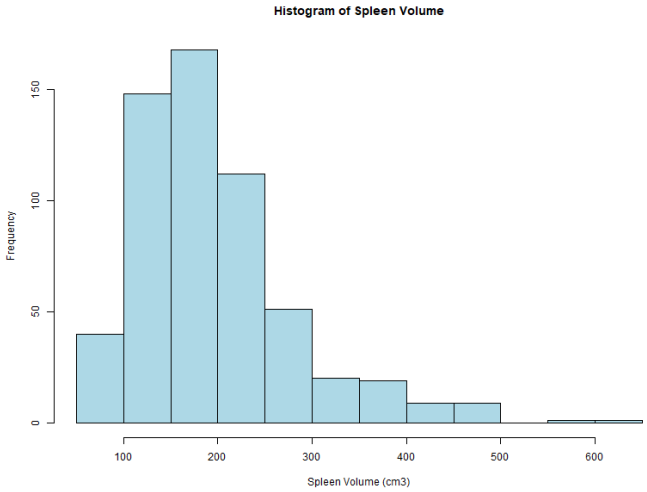
Supplementary Table S4 Data distribution for the prediction model

Variable	n	mean	sd	median	iqr	minimum	maximum
Age (year)	578	54.1	13.7	56.0	20.0	19.0	89.0
Volume (cm^3)	578	195.8	85	177.9	93.7	51.1	644.4
Height (cm)	578	166.8	7.8	167.0	12.0	148	189.0
Weight (kg)	578	68.0	11.9	66.0	15.0	42.0	115.0
BMI (kg/m^2)	578	24.4	3.3	24.2	4.4	14.9	39.6
BSA (m^2)	578	1.9	0.2	1.8	0.2	1.5	2.6

SD, standard deviation; BMI, body mass index; BSA, body surface area; SV, spleen volume; IQR, interquartile range.

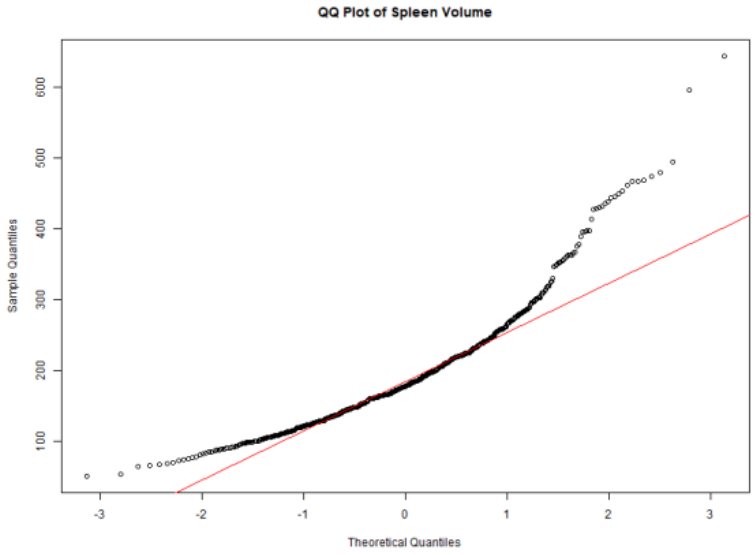
1.2 Spleen volume distribution and normality assessment

Spleen volume was significantly non-normal (Shapiro-Wilk W = 0.899, p < 0.001), with substantial right skew (skewness = 1.45) and heavy tails (kurtosis = 6.04). The distribution is markedly right-skewed and leptokurtic, indicating that most values cluster at the lower end with a long tail of higher values (Supplementary Figures S1 and S2). This non-normal distribution may warrant non-parametric analyses or data transformation for parametric modeling.



Supplementary Figure S1. Histogram of spleen volume distribution.The figure shows the distribution of spleen volumes (in cm3) across the study cohort. The histogram exhibits a clear right-skewed (positive skew) distribution, with the majority of measurements concentrated between 100-200 cm3. Frequencies peak at approximately 200 cm3 and gradually decline as volume increases, demonstrating a positively skewed, non-normal distribution consistent with quantitative normality testing results.

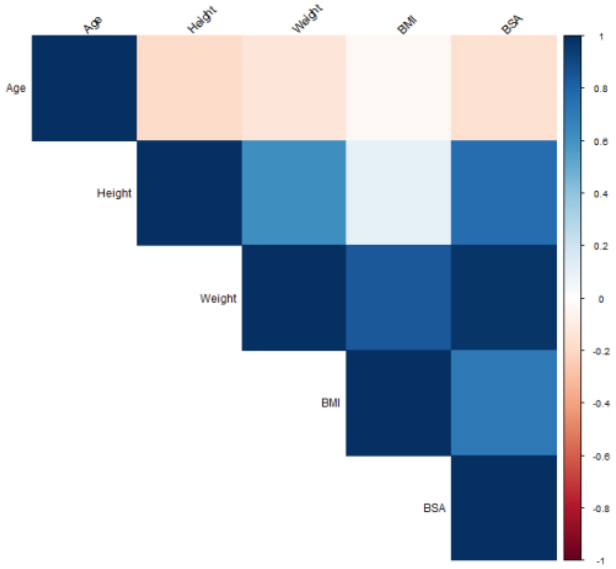
Supplementary Figure S2. Quantile-quantile plot for spleen volume distribution.The plot compares the sample quantiles of spleen volume (y-axis) against theoretical quantiles of a standard normal distribution (x-axis). Data points (black) that deviate substantially from the red diagonal reference line (y = x) indicate departure from normality. The pronounced upward curvature at higher theoretical quantiles (right side) demonstrates heavy right-skewness, with observed spleen volumes substantially exceeding theoretical normal distribution values. This pattern confirms the previously identified non-normal distribution with positive skew.



1.3 Correlation matrix analysis

Age demonstrated negligible correlations with anthropometric measures (r = -0.186 to -0.161).

Height showed moderate positive correlation with Weight (r = 0.619) and BSA (r = 0.767), but minimal correlation with BMI (r = 0.100). Critical multicollinearity concerns emerged between Weight and BSA (r = 0.979) and Weight and BMI (r = 0.840), indicating near-perfect linear relationships. BSA also showed strong correlation with BMI (r = 0.712). The near-unity correlation between Weight and BSA (r = 0.979) suggests these variables contain redundant information, precluding their simultaneous inclusion in regression models (Supplementary Table S5 and Supplementary Figure S3).



Supplementary Figure S3. Correlation heatmap of candidate predictor variables.The heatmap visualizes the Pearson correlation coefficients between Age, Height, Weight, BMI, and BSA. Positive correlations are shown in blue, with color intensity proportional to the strength of the relationship, while negative correlations are shown in red. The diagonal (white squares) represents perfect positive autocorrelation (r = 1) for each variable with itself. High correlations are observed between Weight and BSA (r = 0.979) and between Height and BSA (r = 0.767), highlighting potential multicollinearity among anthropometric predictors. Age shows weak negative correlations with the other measures.

Supplementary Table S5 Correlation coefficients across the predict variables

	Age		Body height		Body weight		BMI		BSA	
	Correlation coefficient	*P* value	Correlation coefficient	*P* value	Correlation coefficient	*P* value	Correlation coefficient	*P* value	Correlation coefficient	*P* value
Age	NA	NA	-0.186	<0.001	-0.137	0.001	-0.039	0.353	-0.161	<0.001
Height	-0.186	<0.001	NA	NA	0.619	<0.001	0.100	0.016	0.767	<0.001
Weight	-0.137	0.001	0.619	<0.001	NA	NA	0.840	<0.001	0.979	<0.001
BMI	-0.039	0.353	0.100	0.016	0.840	<0.001	NA	NA	0.712	<0.001
BSA	-0.161	<0.001	0.767	<0.001	0.979	<0.001	0.712	<0.001	NA	NA

BMI: body mass index, BSA: body surface area

2 Original linear regression model

2.1 Variable selection process

The stepwise algorithm selected a two-predictor model containing BSA and Age. BSA was the strongest predictor, providing the largest initial AIC reduction (from 5137.3 to 5001.4). The addition of age further improved the model (AIC = 4981.0), while no other variables yielded additional AIC improvement. The final model (Volume ~ BSA + Age) demonstrated significant effects for both BSA ($\beta = 202.97$, $p < 0.001$) and Age ($\beta = -1.09$, $p < 0.001$), explaining 24.2% of spleen volume variance. Body height, body weight, and BMI provided no additional predictive value beyond the BSA + Age combination.

2.2 Original model coefficients and performance

The final regression model estimated spleen volume as: Volume = -120.750 + 202.966(BSA) - 1.090(Age). BSA showed a strong positive association ($\beta = 202.966$, 95% CI: 169.078 to 236.853, $p < 0.001$), with each 1 m^2 increase in BSA associated with approximately 203 mL greater spleen volume. Age demonstrated a modest negative association ($\beta = -1.090$, 95% CI: -1.540 to -0.640, $p < 0.001$), with each additional year associated with approximately 1.1 mL lower spleen volume. The model intercept (-120.750) represents the predicted spleen volume for an individual with zero BSA and zero age, though this extrapolation lacks clinical meaning.

Supplementary Table S6 Linear regression coefficients of the original model

	Coefficient	Lower CI	Higher CI	Std Error	t value	P value
(Intercept)	-120.75	-191.819	-49.68	36.184	-3.337	0.001
BSA	202.966	169.078	236.853	17.254	11.764	<0.001
Age	-1.090	-1.540	-0.640	0.229	-4.760	<0.001

BSA, body surface area; SE, standard error; CI, confidence interval.

.3 Model diagnostics and transformation

3.1 Goodness-of-Fit assessment

The model explained 24.2% of spleen volume variance ($R^2 = 0.242$), with adjusted $R^2 = 0.240$ indicating minimal penalty for the two predictors. The overall F-test was highly significant ($F = 91.879$, $p < 0.001$), rejecting the null hypothesis and confirming that at least one predictor contributes significantly to explaining spleen volume. While the model demonstrates statistical significance, the moderate R^2 value suggests additional unmeasured factors influence spleen volume beyond BSA and Age.

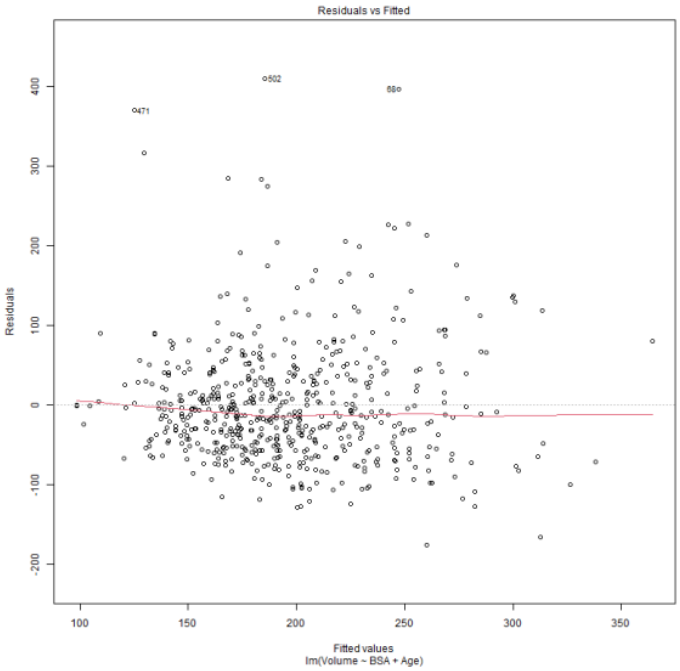
3.2 Multicollinearity assessment

Both predictors in the final model demonstrated negligible multicollinearity: BSA (VIF = 1.026) and Age (VIF = 1.026). Both values are well below the conventional threshold of 5, indicating that the correlation between BSA and Age is minimal and does not inflate coefficient variance. The near-identical VIF values for both predictors reflect their symmetrical contribution to the model's information matrix. This confirms that despite initial concerns about multicollinearity among

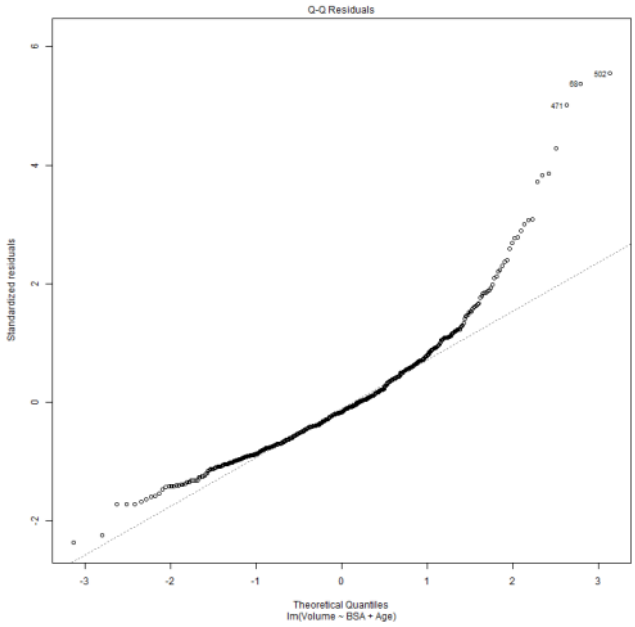
anthropometric measures, the stepwise-selected two-predictor model (BSA + Age) avoids problematic collinearity.

3.3 Diagnostic testing for linear regression assumptions

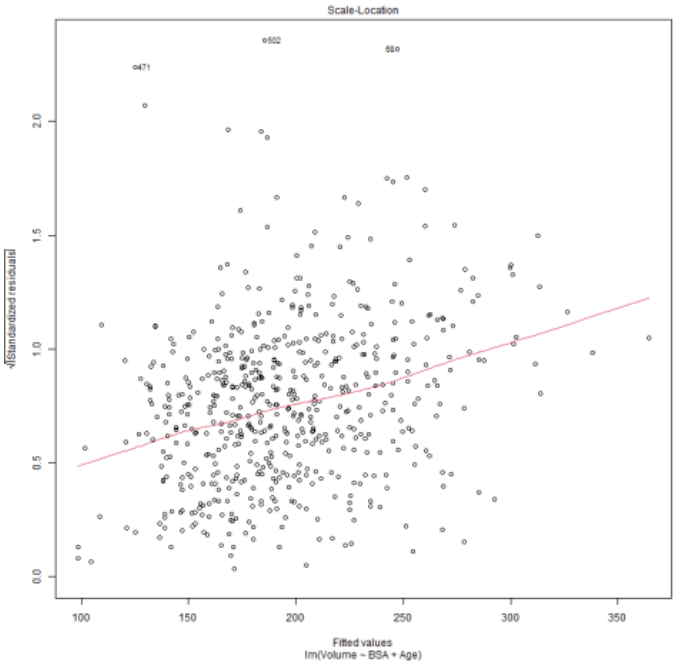
The regression assumptions of normality and homoscedasticity were violated, while the independence assumption was satisfied. Residuals were significantly non-normal (KS test: D = 0.102, p < 0.001), indicating deviation from a normal distribution. Significant heteroscedasticity was detected (Breusch-Pagan test: BP = 8.765, df = 2, p = 0.012), suggesting residual variance changes with predicted values. However, no significant autocorrelation was found (Durbin-Watson statistic = 1.996, p = 0.97), confirming residual independence. The violations of normality and homoscedasticity may affect the reliability of p-values and confidence intervals, suggesting the need for robust standard errors or alternative modeling approaches (Supplementary Figures S4 to S7).



Supplementary Figure S4. Diagnostic plot: Residuals versus fitted values for the initial linear model (Volume ~ BSA + Age). The scatterplot displays the residuals (vertical axis; difference between observed and predicted values) against the model's fitted (predicted) values (horizontal axis). A horizontal red dashed line indicates the zero residual reference. A well-behaving model would show a random scatter of points around this line. The observed pattern—specifically, the widening vertical spread of residuals as fitted values increase and the presence of several extreme outliers (e.g., observations 471, 502, 680)—visually confirms the formal statistical tests, indicating violations of the homoscedasticity (constant variance) and normality assumptions. This diagnostic justifies the subsequent use of a log-transformed response variable.

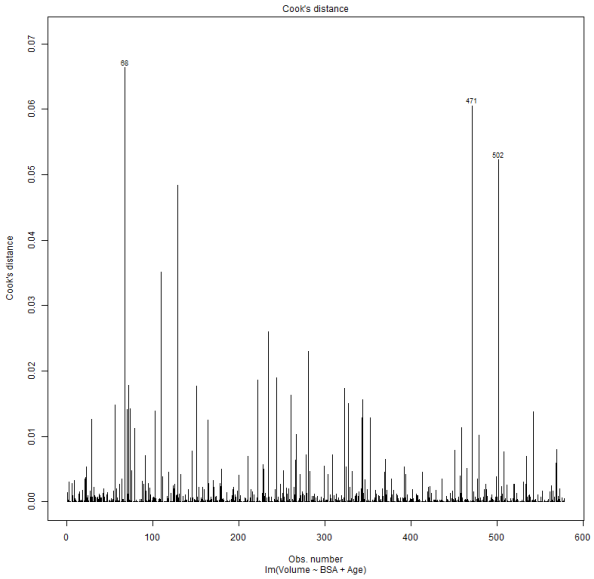


Supplementary Figure S5. Normal quantile-quantile (Q-Q) plot of residuals for the initial linear model (Volume ~ BSA + Age). The plot compares the quantiles of the standardized residuals (y-axis) against the theoretical quantiles of a standard normal distribution (x-axis). In a well-specified model with normally distributed errors, points would closely follow the red diagonal reference line. The pronounced upward curvature in the upper tail (e.g., points labeled 502, 471, 68) and systematic deviations at both tails indicate heavy right-skewness and excess kurtosis, confirming the statistical rejection of normality by formal tests. The extreme positive residuals (above the line) correspond to observations where spleen volume is substantially underpredicted by the model, highlighting influential outliers that violate the normality assumption.



Supplementary Figure S6. Scale-location plot (spread-level plot) for the initial linear model (Volume ~ BSA + Age). The plot displays the square root of the absolute standardized residuals (vertical axis; √|Standardized residuals|) against the fitted (predicted) values (horizontal axis). The red smoothing line (loess curve) aids in visualizing trends in residual spread. A horizontal trend indicates homoscedasticity (constant variance), while an upward or downward slope suggests heteroscedasticity. The observed upward trend in the smoothing line, particularly beyond fitted values of 200, indicates that residual variance increases with larger predicted spleen volumes, confirming the presence of heteroscedasticity detected by the Breusch-Pagan test. Outliers labeled 502, 471, and 680 correspond to observations with high leverage or influence, further highlighting the violation of the homoscedasticity assumption.

Supplementary Figure S7. Influence diagnostic plot: Cook's distance for the initial linear model (Volume ~ BSA + Age). The plot displays Cook's distance (vertical axis) for each observation



(indexed sequentially on the horizontal axis). Cook's distance quantifies the influence of each observation on the overall regression estimates. The plot confirms the presence of strong influential points, which should be examined for data integrity and may warrant robust regression techniques or sensitivity analysis.

.4 Final log-transformed model

4.1 Model coefficients with 95% CI

In the original volume scale, each 1 m² increase in BSA multiplied expected spleen volume by 2.68 (95% CI: 2.29 to 3.14), a substantial proportional increase. Each additional year of age multiplied expected volume by 0.994 (95% CI: 0.992 to 0.996), corresponding to a 0.6% annual decrease. The intercept, 40.79 mL (95% CI: 29.23 to 56.93), represents the geometric mean spleen volume for an individual with BSA = 0 and Age = 0, though this remains a non-clinical reference. All effects were highly significant ($p < 0.001$), with the wide confidence interval for the BSA effect reflecting its strong but variable proportional relationship with spleen volume (Supplementary Table S7).

Supplementary Table S7 Linear regression coefficients of the log-transformed model

	Coefficient	Lower CI	Higher CI	Std Error	t value	P value
(Intercept)	40.79	29.227	56.927	0.17	21.851	<0.001
BSA	2.682	2.288	3.144	0.081	12.192	<0.001
Age	0.994	0.992	0.996	0.001	-5.858	<0.001

Abbreviations: BSA, body surface area; SE, standard error; CI, confidence interval.

Model Equation in Log Scale:

$\log(\text{Spleen Volume}) = 3.708 + 0.987 \times \text{BSA} - 0.00629 \times \text{Age} + \epsilon$

where:

log(Spleen Volume) is the natural logarithm of spleen volume (mL);

BSA is body surface area (m^2);

Age is in years;

ϵ is the random error term, assumed to be normally distributed with mean 0

Equivalent Exponential Form (Back-Transformed):

$\text{Spleen Volume} = e^{3.708} \times e^{0.987 \times \text{BSA}} \times e^{-0.00629 \times \text{Age}} \times e^{\epsilon}$

$\text{Spleen Volume} = 40.79 \times (2.682)^{\text{BSA}} \times (0.994)^{\text{Age}} \times e^{\epsilon}$

Interpretation of Coefficients:

Intercept (3.708): When BSA = 0 and Age = 0, the predicted log(Volume) is 3.708

BSA coefficient (0.987): Each 1 m^2 increase in BSA multiplies expected spleen volume by $e^{0.987}=2.682$(168.2% increase)

Age coefficient (-0.00629): Each additional year multiplies expected spleen volume by $e^{-0.00629}=0.994$(0.6% decrease)

4.2 Goodness-of-Fit assessment of the log-transformed model

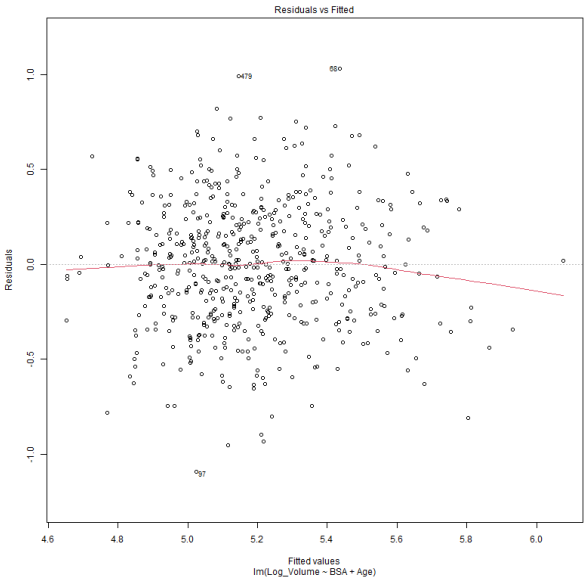
The log-transformed model demonstrated a higher R^2 on the log scale, suggesting improved linearity and variance stabilization, though R^2 values across different outcome scales are not directly comparable. The overall F-test was highly significant ($F = 105.669$, $p < 0.001$), confirming that BSA and Age jointly explain a statistically significant proportion of variance in spleen volume. The increased R^2 suggests the log-transformation better captures the relationship between predictors and spleen volume, though substantial unexplained variance remains.

4.3 Multicollinearity assessment

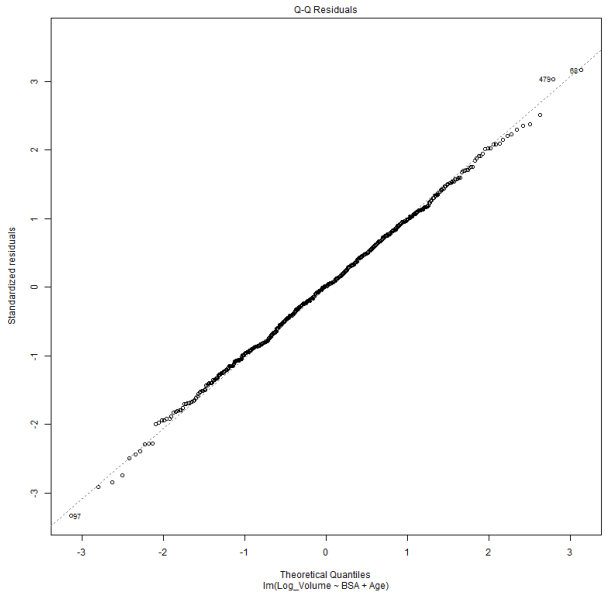
Both predictors in the log-transformed linear model demonstrated negligible multicollinearity: BSA (VIF = 1.026) and Age (VIF = 1.026). The identical VIF values reflect the symmetrical relationship between these two uncorrelated predictors in the model. Both values are substantially below the threshold of 5, confirming that the correlation between BSA and Age is minimal and does not compromise the stability or interpretability of the coefficient estimates. The transformation of the response variable to the log scale did not introduce or exacerbate any collinearity issues among the predictors.

4.4 Diagnostic testing for linear regression assumptions of the log-transformed model

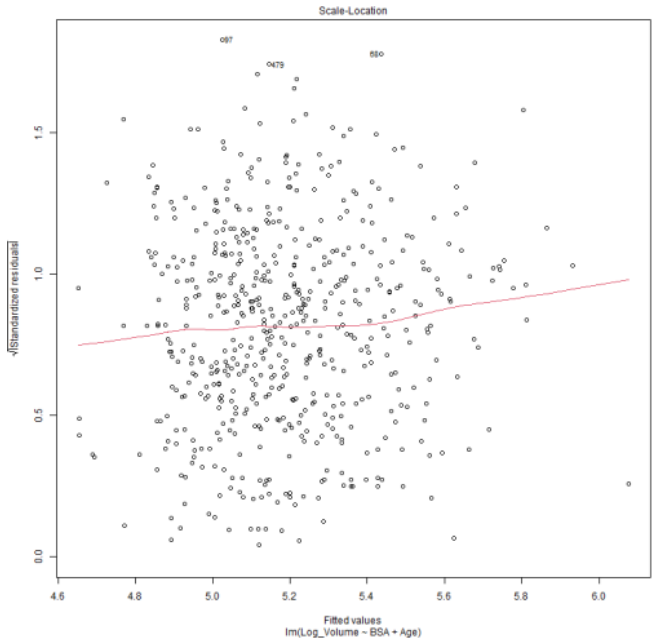
The log-transformation successfully addressed the assumption violations present in the original model. Residuals were not significantly different from a normal distribution (KS test: D = 0.023, p = 0.912). Heteroscedasticity was no longer statistically significant at the 0.05 level (Breusch-Pagan test: BP = 4.997, df = 2, p = 0.082), though marginal evidence remains. Residuals showed no significant autocorrelation (Durbin-Watson statistic = 1.919, p = 0.37), satisfying the independence assumption. This demonstrates that the log-transformation effectively normalized the residual distribution and stabilized the variance, making the model's statistical inference more reliable (Supplementary Figures S8 to S11).



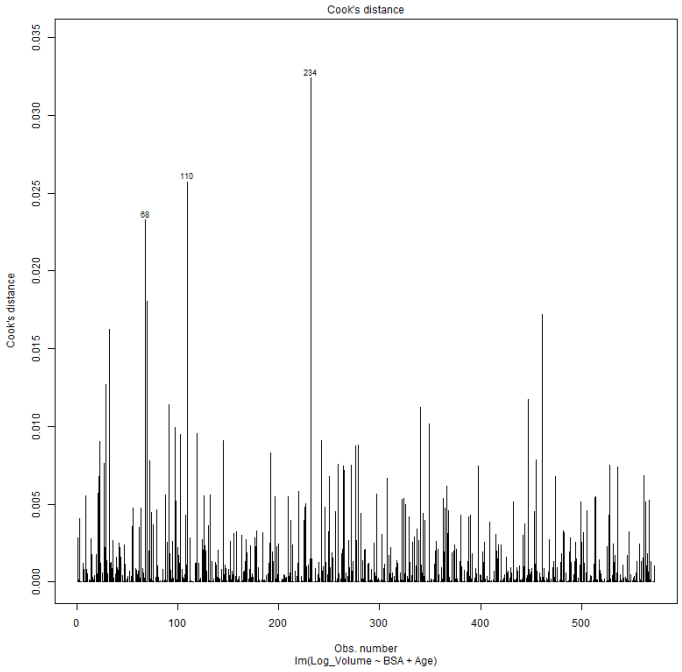
Supplementary Figure S8. Residuals versus fitted values for the log-transformed linear regression model (Log_Volume ~ BSA + Age). The plot displays the residuals (vertical axis; difference between observed and predicted log-transformed values) against the model's fitted values (horizontal axis; predicted log(Volume)). The pink curve represents a smoothed loess fit, and a horizontal reference line at residual = 0 is implied. The points are generally distributed randomly around the center, though a subtle, shallow U-shaped pattern in the loess curve suggests minimal residual structure. Three flagged observations (97, 479, 680) are potential outliers. The lack of a pronounced funnel shape or systematic trend indicates that the log transformation has largely addressed the heteroscedasticity evident in the original model (Supplementary Figure S4), aligning with the non-significant Breusch-Pagan test result.



Supplementary Figure S9. Normal quantile-quantile (Q-Q) plot of residuals for the log-transformed linear regression model (lm(Log_Volume ~ BSA + Age)). The plot compares the standardized residuals of the model (y-axis) against the theoretical quantiles of a standard normal distribution (x-axis). The gray diagonal reference line represents perfect normality. The vast majority of data points closely follow the reference line, indicating that the residuals of the log-transformed model are approximately normally distributed—a marked improvement over the original model (Supplementary Figure S5). Minor deviations observed in the upper and lower tails are within acceptable limits for large-sample inference, supporting the formal test result (Kolmogorov-Smirnov p = 0.912) that the normality assumption is satisfied.



Supplementary Figure S10. Scale-location plot (spread-level plot) for the log-transformed linear regression model (lm(Log_Volume ~ BSA + Age)). The plot displays the square root of the absolute standardized residuals (√|Standardized residuals|, vertical axis) against the fitted values (horizontal axis) to assess homoscedasticity. The red smooth curve (loess) shows a nearly horizontal trend across the range of fitted values, with only minimal localized fluctuation, indicating that the variance of the residuals is approximately constant. The labeled points (0.97, 0.97, 0.90) correspond to three observations that slightly deviate from the main pattern but do not indicate systematic heteroscedasticity. The overall flatness of the smooth line supports the non-significant result of the Breusch-Pagan test (p = 0.082), confirming that the log transformation successfully stabilized the variance of the residuals.



Supplementary Figure S11. Residuals versus leverage plot for the log-transformed linear regression model (Log_Volume ~ BSA + Age). The plot displays standardized residuals (vertical axis) against leverage (horizontal axis), a measure of how unusual each observation's predictor values are. The majority of points cluster in the lower left, confirming that most observations have low leverage and well-behaved residuals. This plot helps identify points that are both outliers in the response (high residuals) and outliers in the predictor space (high leverage), which warrant careful review.

5 Internal validation

5.1 Bootstrap validation results

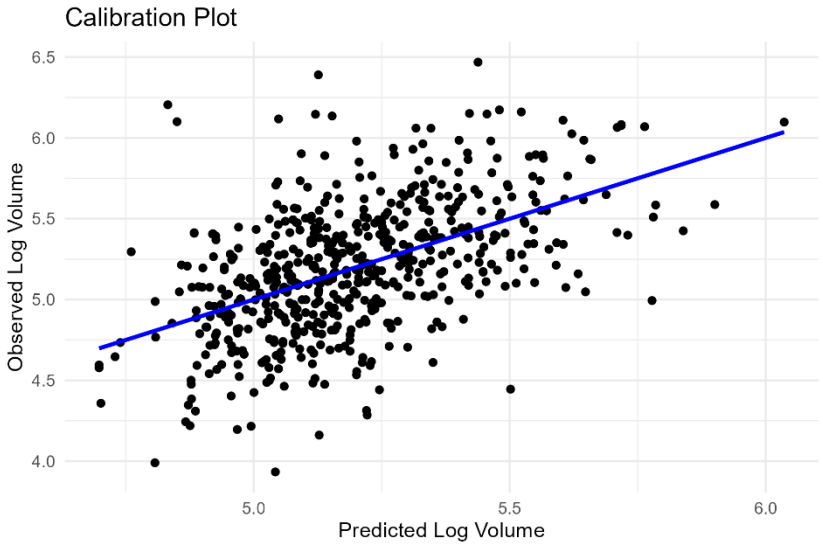
The apparent R^2 (0.269) and RMSE (0.347) were minimally optimistic for the log-transformed linear regression model (Log_Volume ~ BSA + Age). After bootstrap correction, the optimism-adjusted R^2 was 0.261, a reduction of only 0.008, and the adjusted RMSE was 0.351, an increase of 0.004. The optimism estimates themselves were small and variable: the 95% CI for R^2

optimism ranged from -0.136 to 0.152, and for RMSE optimism from -0.057 to 0.053, both intervals containing zero. The narrow shift between apparent and corrected metrics, combined with the wide optimism CIs spanning zero, suggests that the model's in-sample performance is a reasonably unbiased estimate of its true predictive ability, with little evidence of substantial overfitting.

5.2 Calibration assessment

The calibration plot demonstrated close agreement between observed and predicted log-transformed spleen volumes across the range of fitted values (Supplementary Figure S12). The estimated calibration slope was close to 1, and the calibration intercept was close to 0, indicating no substantial systematic over- or under-prediction on the log scale.

As expected for an in-sample assessment of a linear regression model, these calibration metrics reflect the fitted model's internal consistency and should not be interpreted as evidence of external calibration performance. The bootstrap validation results, which demonstrated minimal optimism in model performance, support the stability of these calibration findings.



Supplementary Figure S12. Calibration plot for the log-transformed linear regression model (Log_Volume ~ BSA + Age). The plot displays the observed log-transformed spleen volume (Observed Log Volume) on the y-axis against the model-predicted log-transformed spleen volume (Predicted Log Volume) on the x-axis. Each point represents an individual observation (n = 578). The blue line corresponds to the regression of observed on predicted values, with a slope of 1.0 and an intercept of 0.0, indicating perfect in-sample calibration. The majority of points cluster around the identity line, with predicted values ranging approximately from 5.0 to 6.0 and observed values from 4.5 to 6.0. The spread of points about the line reflects the residual variation not explained by the model ($R^2 = 0.269$). The tight alignment of the blue line with the identity line confirms that the model does not exhibit systematic over- or under-prediction across the range of fitted values, though the calibration is inherently optimistic because the same data were used for both fitting and evaluation.

6 Model stability and generalizability

The model demonstrated excellent stability metrics:

EPV ratio = 289 (n = 578, p = 2), far exceeding the recommended minimum of 10–20 for linear

regression

Shrinkage factor = 1.0 (95% CI: 0.995–1.006), indicating no evidence of overfitting

Bootstrap CIs confirmed precision, with the shrinkage factor's lower 95% bound (0.995) still exceeding the 0.9 threshold.

These results confirm the model is not overfit to the sample data and can be expected to perform similarly in new datasets. The extremely high EPV ratio provides additional confidence in the reliability of the coefficient estimates.

7 Sensitivity analyses

The full model (BSA + Age) demonstrated the best overall performance across all metrics (Supplementary Table S8).

While BSA alone explained most of the variance (R^2 = 0.225), adding Age provided a statistically and practically significant improvement (ΔR^2 = +0.044). The Age-only model performed substantially worse than both alternatives. All models showed good calibration (mean residuals ≈ 0), but the full model had the tightest residual distribution (SD = 0.347). These results justify retaining both predictors in the final model despite their moderate correlation, as the combined model offers superior predictive accuracy without evidence of overfitting.

Supplementary Table S8 Sensitivity analysis of the log-transformed linear regression model

Model	AIC	Mean_Residual	SD_Residual	R_squared	MAE	RMSE
BSA only	456.042	-8.22237E-18	0.357	0.225	0.279	0.357
Age only	555.434	1.95383E-17	0.390	0.08	0.307	0.389
BSA + Age	424.534	-4.73119E-19	0.347	0.269	0.269	0.347

RMSE, root mean square error; R^2, coefficient of determination.

8 Interaction and subgroup analyses

8.1 Interaction analysis by sex

The interaction model did not provide a statistically significant improvement over the additive model (ANOVA F = 1.528, df = 3, p = 0.206). Neither interaction term was significant when tested individually:

BSA × Sex: F = 0.300, p = 0.584 (coefficient = -0.116, t = -0.548, p = 0.584)

Age × Sex: F = 1.178, p = 0.278 (coefficient = +0.002, t = 1.085, p = 0.278)

The main effects of BSA ($p < 0.001$) and Age ($p < 0.001$) remained significant, while the Sex main effect was non-significant (p = 0.733). The interaction model explained only marginally more variance (R^2 = 0.275 vs. 0.269 in the additive model). These results suggest that the relationships between BSA/Age and spleen volume are consistent across sexes, justifying the use of the simpler additive model without sex interactions.

8.2 Sex-specific models

The stratified model was not a statistically significant improvement over the pooled model (F = 1.528, df = 3, p = 0.206). This confirms that neither the association of BSA with spleen volume nor that of Age with spleen volume differs significantly by sex, and that the baseline spleen volume (intercept) does not differ by sex after adjusting for BSA and Age. These findings are consistent

with the non-significant interaction terms in the multiplicative model and support the use of the simpler pooled model without sex stratification. The minimal difference in residual sums of squares (ΔRSS = 0.553) relative to the degrees of freedom added (Δdf = 3) further confirms that sex-specific modeling does not meaningfully improve fit (Supplementary Tables S9 and S10).

Supplementary Table S9. Coefficients of log-transformed model in subgroups of sex

	Male	Female	Pooled
Intercept	3.969	3.824	3.708
BSA	0.841	0.958	0.987
Age	-0.005	-0.008	-0.006

BSA, body surface area; SE, standard error; CI, confidence interval.

Supplementary Table S10 Performance of subgroups models by Sex

	Mean_residual	Sd_residual	R2	RMSE	Cal_slope.	Cal_intercept
Male	7.88803E-18	0.371	0.174	0.37	0.995	0.995
Female	-8.44724E-18	0.319	0.204	0.318	0.996	0.996
Pooled	-4.73119E-19	0.347	0.269	0.347	0.996	0.996

RMSE, root mean square error; R^2, coefficient of determination.