

ORIGINAL RESEARCH ARTICLE

Development of a machine learning-based risk prediction model for lymphoma using data from the 2023 National Health Interview Survey

Yuying Guo¹, Shuling Hou¹, Xuejing Yang¹, and Dong Song¹

Cancer Center, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Third Hospital of Shanxi Medical University, Tongji Shanxi Hospital, Taiyuan, China

Abstract

Lymphoma is a malignant tumor that originates from the lymphatic system. This study aims to develop a risk prediction model for lymphoma using lymphoma cases extracted from the 2023 National Health Interview Survey database. The χ^2 test was used to examine differences in seven variables between groups. Subsequently, hyperparameter combinations were randomly tested for optimization. Four machine learning models were constructed after obtaining the optimal hyperparameters. Receiver operating characteristic analysis was employed to evaluate the four machine learning models and select the most effective model. The relative importance of the seven variables was ranked using the machine learning algorithm. The data from the baseline characteristic table, comprising 3,670 participants, revealed that six variables differed significantly ($p < 0.05$). After hyperparameter tuning, the extreme gradient boosting (XGBoost) model achieved the optimal parameter combination, with a receiver operating characteristic value of 0.8567. The XGBoost model also demonstrated the highest area under the curve value (0.8161), validating it as the best-performing model, with a sensitivity of 0.9916, a specificity of 0.1188, a precision of 0.9086, an F1 score of 0.9483, and an accuracy of 0.9028. Finally, the machine learning model found that obesity (body mass index ≥ 30 kg/m²) exerted the greatest predictive contribution to lymphoma, with a relative importance score of 100. This study analyzed seven variables associated with lymphoma occurrence and developed a risk prediction model for lymphoma, providing valuable insights into the treatment of this disease.

*Corresponding authors:

Shuling Hou
(houshuling@sx bqeh.com.cn)
Xuejing Yang
(yangxuejing@sx bqeh.com.cn)
Dong Song
(songdong@sx bqeh.com.cn)

Citation: Guo Y, Hou S, Yang X, Song D. Development of a machine learning-based risk prediction model for lymphoma using data from the 2023 National Health Interview Survey. *Cancer Plus*. 2026;8(3):025450072. doi: 10.36922/CP025450072

Received: November 7, 2025

Revised: December 8, 2025

Accepted: December 19, 2025

Published online: December 24, 2025

Copyright: © 2025 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Lymphoma; Risk prediction model; National Health Interview Survey; Baseline characteristic; Machine learning

1. Introduction

Lymphoma, the sixth most common type of cancer worldwide,¹ is a group of malignant tumors that originate from lymph nodes and extralymphatic tissues.² It can be categorized into two types: Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL). Approximately 5–10% of lymphomas are HL, which is divided into classical and nodular lymphocyte-predominant subtypes. In contrast, NHL is further divided into B-cell, T-cell, and natural killer (NK) cell subtypes.³ In 2020, a survey reported 544,352 new NHL cases globally, ranking it 13th among new malignant tumor cases worldwide,

with 55.9% of cases in males and 44.1% in females.⁴ The overall mortality rate of NHL was reported to be 28.1%. In addition, a total of 83,087 HL cases were reported (males: 59.1%; females: 40.9%), with an overall mortality rate of 47.7%. Recently, targeted therapies for lymphoma have made significant advancements. However, some patients responded poorly due to drug resistance, resulting in high rates of treatment failure and relapse.⁵ Hematopoietic stem cell transplantation remains an important treatment for relapsed/refractory lymphoma. However, outcomes are still constrained by the source and quality of hematopoietic stem cells, as well as their post-transplant developmental and functional status.⁶ Therefore, understanding the various factors contributing to lymphoma development and establishing a risk prediction model are crucial for aiding the prevention and early diagnosis of lymphoma.

The National Health Interview Survey (NHIS) aims to collect data on the health status, medical history, use of medical services, and health behaviors of the United States (U.S.) population and provides basic data for health policy formulation, health research, and health service planning. A multi-stage stratified sampling method was used to select families and individuals from U.S. states and counties as participants, ensuring the sample is representative of the U.S. population across different regions and socioeconomic backgrounds.⁷ It offers strong national representativeness, comprehensive health information, and a broad range of variables.^{7,8} At present, an increasing number of researchers have used the NHIS database to conduct various studies, providing important references for people's health. For example, Mbous *et al.*⁷ conducted a study based on the NHIS database and found that education, family income-to-poverty ratio, body mass index (BMI), number of chronic conditions, alcohol use, and general health as key factors accounting for the physical activity (PA) disparity between White and Black or Multiple/Mixed-race cancer survivors. These findings could help inform behavioral PA interventions to improve their design and targeting toward different racial groups of cancer survivors.

Based on the NHIS data, the current study identified factors associated with lymphoma occurrence, such as age and gender, and used these lymphoma-associated variables to evaluate four machine learning models: random forest (RF), extreme gradient boosting (XGBoost), support vector machine (SVM), and least absolute shrinkage and selection operator (LASSO), and compared their performance to identify the best-performing model. Finally, the relative importance scores of lymphoma-related risk variables were ranked, providing a solid foundation for an in-depth understanding of lymphoma pathogenesis and the development of new treatment strategies.

2. Materials and methods

2.1. Data collection

Data were obtained from the NHIS. The NHIS, which began in 1957 and is a continuing survey, documents the prevalence, distribution, and effects of illness and disability in the U.S. population. Our study included adult participants with a history of lymphoma from the 2023 NHIS dataset. The exclusion criteria were (i) unclear lymphoma diagnosis or missing key information and (ii) participants with missing data for other variables (age, gender, race, BMI, breast cancer, hypertension, and health status). A total of 3,670 participants, comprising 101 patients with lymphoma and 3,569 individuals without lymphoma, were included. The flowchart is presented in Figure 1.

2.2. Definition of variables

The corresponding numbers and grouping information for each variable were as follows.

For age (AGEP_A), all participants were over 18 years old, and they were divided into four age ranges: 18–24 years, 25–44 years, 45–64 years, and 65 years and older. For gender (SEX), they were divided into male and female, and for race (HISPALLP_A), they were divided into non-Hispanic White, Hispanic, non-Hispanic Black/African American, Non-Hispanic American Indian or Alaska Native (AIAN), non-Hispanic AIAN of any other group, non-Hispanic Asian, and other single and multiple races. For BMI (BMICAT_A), participants were categorized into underweight, healthy weight, overweight, and obese groups. For hypertension (HYPERV_A), participants who responded “yes” to the question “Has a doctor or other health professional ever told you that you have hypertension?” were classified as having hypertension (Yes), while the others were classified as non-hypertensive

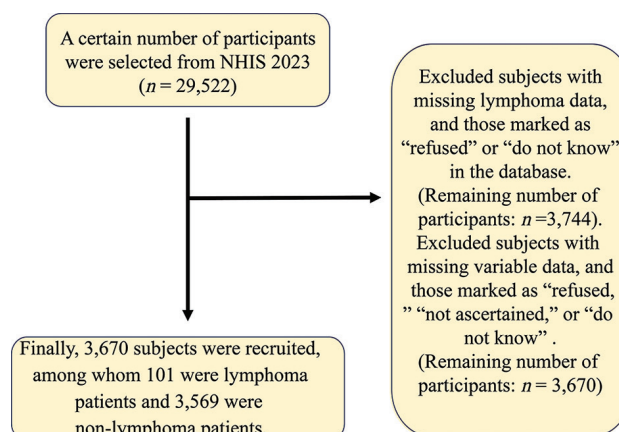


Figure 1. Flowchart of inclusion and exclusion of subjects
Abbreviation: NHIS: National Health Interview Survey.

(No), while for breast cancer (BREASCAN_A), participants who recorded “mentioned” to the question “Has a doctor or other health professional ever told you that you have breast cancer?” were classified as breast cancer (Yes), while the others were classified as non-breast cancer (No). For health status (PHSTAT_A), they were divided into excellent, very good, good, fair, and poor.

2.3. Outcome definition

The 2023 NHIS database was used to extract data from the adult questionnaire module, and the variable LYMPHCAN_A was queried. The participants were asked, “Sample adults 18+ who were ever told they had cancer?” Participants who answered “Mentioned” to the lymphoma-specific query were classified as lymphoma cases, while those who answered “Not Mentioned” were classified as controls.

2.4. Statistical analysis

A baseline table was constructed using the R package “tableone” (v0.13.2, R Foundation for Statistical Computing, Austria)⁹ based on the seven variables. Categorical variables were shown as percentages. Group differences in continuous variables were quantified using Cohen’s d effect size, calculated through the cohen.d() function from the R package “effectsize”¹⁰ (v0.8.5), with 95% confidence intervals (CIs) reported. The strength of association for categorical variables was assessed using Cramer’s V effect size, computed through the cramersV() function in the R package “rcompanion”¹¹ (v2.5.1); the χ^2 test was performed using the Chi-sq.test() function from the stats package to obtain the χ^2 value, and statistical significance was set at $p < 0.05$. The R package “caret” (v6.0.94)¹² was used to adjust hyperparameters. For all models, the parameters were set as method = cv, number = 5, search = random. A 5-fold cross-validation was adopted. For LASSO and SVM, an additional parameter, search, was set to random, and 10 random searches were conducted to evaluate 10 hyperparameter combinations.

The R package “glmnet” (v4.1.8)⁹ was utilized to conduct LASSO regression analysis on variables with group differences. The regularization parameters, alpha and lambda, control the intensity of regularization and, thus, the model’s complexity.

The R package “xgboost” (v2.0.3.1)⁹ was used to perform an XGBoost analysis of variables with group differences. A parameter grid was constructed by selecting specific parameters. Among them, nrounds is the number of boosting iterations, eta is the learning rate, which refers to the weight reduction coefficient of each tree, and Max_depth refers to the uppermost limit of depth that a solitary

tree can reach. Gamma represents the minimum reduction in loss required when splitting a node. Colsample_bytree denotes the fraction of features that are randomly picked for each tree. Min_child_weight is defined as the least sum total of sample weights across the child nodes. Subsample, on the other hand, is the proportion of training samples that are randomly selected for each individual tree.

The R package “e1071” (v1.7.14)⁹ was used to perform SVM analysis on variables with group differences. C is the regularization parameter, and sigma is the width parameter of the radial basis kernel function.

The R package “forestplot” (v3.1.3)⁹ was used to perform RF analysis of variables with group differences. The parameter mtry is a key hyperparameter that controls the number of features randomly selected at each decision tree split in a node.

The parameters with the highest receiver operating characteristic (ROC) values in each model were selected for evaluation. To determine the optimal model, the performances of the four machine learning models after hyperparameter tuning were compared. The R package “caret” was used to conduct ROC analysis to analyze the sensitivity, specificity, precision, F1 score, accuracy, and area under the curve (AUC) of the four machine learning models. The AUC was used to determine the optimal model ($0.6 < \text{AUC} < 1.0$).

To comprehensively evaluate the classification performance of machine learning models, confusion matrices were constructed and relevant metrics calculated. The confusionMatrix() function from the “caret” package was used to generate confusion matrices for each model, distinguishing true positives (TPs), true negatives, false positives (FPs), and false negatives (FNs). Based on these matrices, accuracy, sensitivity (i.e., recall), specificity, precision, and F1 score were computed using the R packages caret and “yardstick” (v 1.2.0).¹³ The R package “pROC”¹⁴ (v1.18.4) was applied to analyze FP and FN ratios, assessing the misclassification patterns of models and their clinical implications in the context of imbalanced datasets.

The R package “binom”^{15,16} (v1.1-1.1), based on the binomial distribution, was used to calculate the positive predictive value, negative predictive value (NPV), and their 95% CIs. Calibration curves were plotted using the R package “rms”¹⁷ (v6.7-0), and calibration quality was quantified by multiple statistics: Brier score (calculated through mean squared error), calibration slope and intercept (derived from logistic regression), and the Hosmer–Lemeshow goodness-of-fit test (from the “ResourceSelection” package¹⁸ [v0.3-6]). Finally, to assess

the predictive importance of the seven variables for lymphoma risk, variable importance scores were calculated using the “caret” package and subsequently ranked in descending order. All statistical analyses were performed using R (version 4.2.2), and $p < 0.05$ was considered statistically significant.

3. Results

3.1. Statistical analysis of baseline characteristics of lymphoma patients

The variables with marked statistical disparities between the group with lymphoma and the non-lymphoma control group included age, race, BMI, breast cancer, hypertension,

and health status ($p < 0.05$). Specifically, in comparison to the control group, the proportion of those aged 65 and above with cancer was higher. Among the participants with lymphoma, the proportions of those who were overweight and obese were 32.7% and 34.7%, respectively, which indicated that those with a BMI ≥ 25 were more likely to develop cancer. In addition, hypertension and breast cancer may have a certain association with the risk of lymphoma or the disease state (Table 1).

3.2. Evaluation of the optimal machine learning model

Results of the four models showed that the LASSO model achieved a relatively good fit when $\alpha = 0.5756$,

Table 1. Baseline table features

Variables	Level	Non-lymphoma, <i>n</i> (%)	Lymphoma, <i>n</i> (%)	<i>p</i>	Effect size	CI	Test statistic, χ^2 (df)
Total		3,569	101				
Age (years)	18–24	10 (0.3)	1 (1.0)	0.043	Cramér's V=0.047	0.023–0.084	χ^2 (3) = 8.169
	25–44	224 (6.3)	10 (9.9)				
	45–64	888 (24.9)	33 (32.7)				
	≥ 65	2,447 (68.6)	57 (56.4)				
Gender	Male	1,542 (43.2)	54 (53.5)	0.051	Cramér's V=0.032	0.003–0.065	χ^2 (1) = 3.8
	Female	2,027 (56.8)	47 (46.5)				
Race	Hispanic	201 (5.6)	17 (16.8)	< 0.001	Cramér's V=0.081	0.059–0.12	χ^2 (6) = 24.221
	Non-Hispanic White	3,030 (84.9)	79 (78.2)				
	Non-Hispanic Black/ African American	226 (6.3)	3 (3.0)				
	Non-Hispanic Asian	58 (1.6)	1 (1.0)				
	Non-Hispanic AIAN	13 (0.4)	0 (0.0)				
	Non-Hispanic AIAN any other group	25 (0.7)	1 (1.0)				
	Other single and multiple races	16 (0.4)	0 (0.0)				
BMI	Underweight	56 (1.6)	5 (5.0)	0.037	Cramér's V=0.048	0.024–0.085	χ^2 (3) = 8.468
	Healthy weight	1,133 (31.7)	28 (27.7)				
	Overweight	1,314 (36.8)	33 (32.7)				
	Obese	1,066 (29.9)	35 (34.7)				
Breast cancer	Yes	728 (20.4)	4 (4.0)	< 0.001	Cramér's V=0.065	0.033–0.098	χ^2 (1) = 15.607
	No	2,841 (79.6)	97 (96.0)				
Hypertension	Yes	2,054 (57.6)	45 (44.6)	0.012	Cramér's V=0.041	0.009–0.074	χ^2 (1) = 6.256
	No	1,515 (42.4)	56 (55.4)				
Health status	Excellent	406 (11.4)	16 (15.8)	0.008	Cramér's V=0.061	0.038–0.099	χ^2 (4) = 13.783
	Very good	1,039 (29.1)	23 (22.8)				
	Good	1,174 (32.9)	27 (26.7)				
	Fair	690 (19.3)	19 (18.8)				
	Poor	260 (7.3)	16 (15.8)				

Abbreviations: AIAN: American Indian or Alaska Native; BMI: Body mass index; CI: Confidence interval; df: Degree of freedom.

lambda = 0.0039, and the ROC curve area was 0.6889 (Table S1). For the XGBoost model, by adjusting hyperparameters, the optimal parameter combination was found to be gamma = 0, eta = 0.1, max_depth = 8, colsample_bytree = 0.8, nrounds = 53, min_child_weight = 1, subsample = 1, yielding an ROC value of 0.8567 (Table S2). When evaluating the SVM model, the ROC values for the linear, radial basis, and polynomial kernels were 0.5448, 0.7210, and 0.6557, respectively. Therefore, the optimal parameters for the radial basis kernel function, with sigma = 0.0087 and C = 83.4374, were selected for the SVM model (Table S3). Subsequently, an RF model was evaluated with mtry = 11, yielding an ROC value of 0.6889 (Table S4). Finally, following hyperparameter tuning, the four machine learning models were evaluated using ROC curve analysis. The XGBoost model had the highest AUC value (Figure 2A). Its sensitivity was 0.9916, specificity was 0.1188, precision was 0.9086, F1 score was 0.9483, and accuracy was 0.9028 (Table S5). Confusion matrix analysis revealed that all models demonstrated excellent predictive performance for the X0 class, with accuracies ranging from 88.8% to 89.8% and the number of correct predictions exceeded 3,500 cases. Among them, the XGBoost model achieved 89.2% accuracy for the X0 class, demonstrating stable recognition ability for the major class (Figure 2B). Further comprehensive evaluation indicated that all models had high sensitivity (0.989–1.000) and NPV (0.989–1.000), demonstrating excellent ability to detect positive cases and to reliably rule out non-diseased individuals (Table S6). Calibration analysis results revealed that the XGBoost model had the lowest Brier score (0.0708) and a Hosmer–Lemeshow statistic consistent with that of the SVM model (37.39), with its calibration curve closest to the ideal reference line, indicating optimal calibration accuracy and reliability (Figure 2C). In summary, the XGBoost model evaluated using differential variables, such as age and BMI, after optimal hyperparameter selection, not only achieved the highest AUC for discriminative ability but also demonstrated the best calibration of predicted probabilities, indicating good effectiveness and reliability in predicting lymphoma risk.

3.3. Importance ranking of the seven variables

Given that lymphoma risk is influenced by multiple factors, the XGBoost model showed that obesity (BMI ≥ 30 kg/m²) had the strongest predictive impact, with an importance score of 100. The non-Hispanic AIAN (other subgroups) category had the lowest importance score (5.83) in the model, indicating that race/ethnicity had a relatively weak predictive effect on lymphoma risk. These findings suggested that developing targeted, precision intervention

strategies focused on obesity and other high-impact factors could help reduce lymphoma risk (Figure 3 and Table S7).

4. Discussion

Lymphoma is a malignant tumor that originates from lymph nodes and extralymphatic tissues.² There are two categories: HL and NHL. Each lymphoma subtype differs in clinical manifestations, pathogenesis, treatment strategies, and prognosis evaluation.^{19,20} Based on the 2023 NHIS database, we conducted a population-level retrospective cohort study to identify risk factors for lymphoma and build a predictive machine learning model.^{21,22}

In this study, data related to lymphoma were extracted from the NHIS database, including seven variables: age, gender, race, BMI, history of breast cancer, history of hypertension, and health status. By constructing and comparing the baseline characteristic table, it was found that the difference in BMI distribution between the two groups was statistically significant ($p=0.037$), suggesting that BMI may be related to the risk of lymphoma or disease status; the increase in age was significantly related to the increased risk of lymphoma, especially in people over 65 years old ($p=0.043$), and the susceptibility of this group was higher. In addition, the history of breast cancer and hypertension may also be related to the risk of lymphoma or disease status. In summary, BMI, age²¹ (especially people over 65), history of breast cancer, and history of hypertension and other multi-dimensional variables were significantly related to the risk of lymphoma, indicating that the occurrence of lymphoma may be affected by a variety of factors.²² The ROC curve graphically depicts the tradeoff between sensitivity and specificity at different test result cutoffs for defining a positive test. The goal for a diagnostic test or prediction rule is to have as many TPs and as few FPs as possible. This corresponds to moving up the ROC curve as much as possible before moving to the right.²³

We combined the ROC analysis results to compare the performance of the four machine learning models after hyperparameter tuning, identifying the optimal model. ROC analysis can evaluate the sensitivity, specificity, precision, F1 score, accuracy, and AUC of different models. AUC is a comprehensive metric for assessing classification performance that considers changes in sensitivity and specificity across all possible thresholds, is not influenced by threshold selection, and provides a more comprehensive reflection of the model's discriminative ability. Specifically, the AUC values for the four models were RF = 0.7081, SVM = 0.7492, XGBoost = 0.8161, and LASSO = 0.6994, indicating that XGBoost is the best model for lymphoma risk prediction. XGBoost is a regularized, tree-based

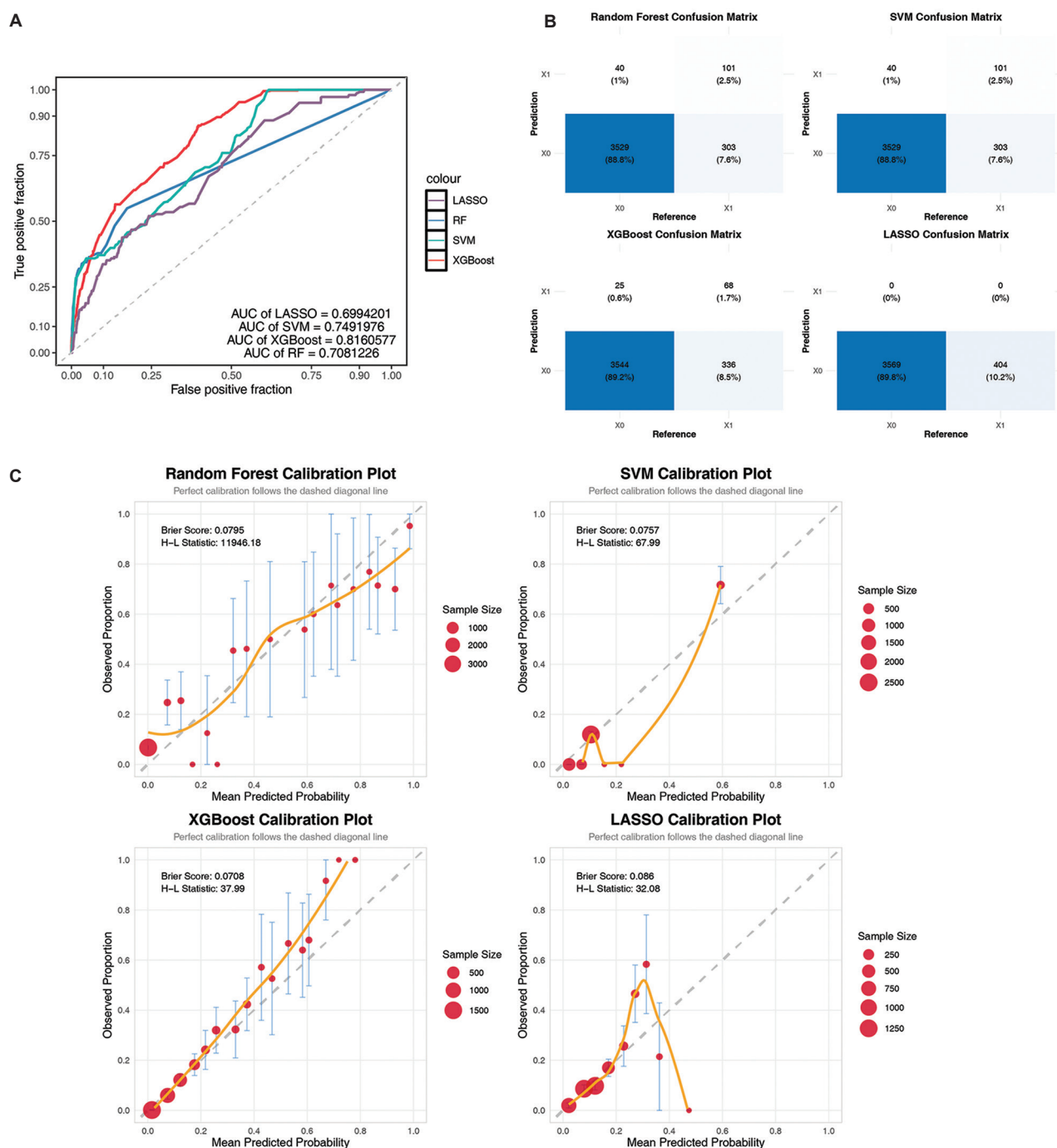


Figure 2. The XGBoost model as the optimal model for lymphoma risk prediction. (A) Receiver operating characteristic curves of the four models. The X-axis represents false positives, the Y-axis represents true positives. (B) Confusion matrix: The X-axis represents the actual true category of the sample (i.e., the real grouping result); the Y-axis represents the model's predicted category. (C) Calibration curve: The X-axis represents the mean predicted probability of the model classifying the sample as a "positive event"; the Y-axis represents the observed proportion of samples that actually experienced the "positive event" within the corresponding predicted probability group.

Abbreviations: AUC: Area under the curve; H-L: Hosmer-Lemeshow; LASSO: Least absolute shrinkage and selection operator; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

boosting machine learning model. The method can turn heuristic rules into highly accurate predictions.²⁴

Previous predictive studies on lymphoma risk have primarily relied on traditional logistic regression or Cox

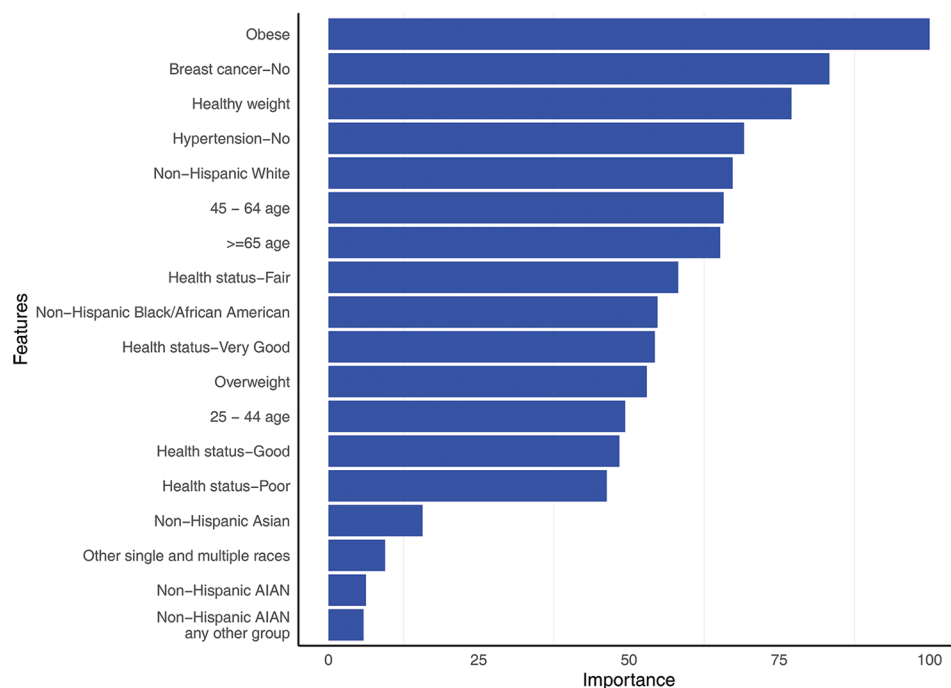


Figure 3. Ranked importance scores of features. Y-axis: Category of different variables; X-axis: Variable importance score
Abbreviation: AIAN: American Indian or Alaska Native.

proportional hazard models, focusing on specific clinical or genetic factors. For example, a prognostic model for diffuse large B-cell lymphoma (DLBCL) based on clinical parameters (such as age, stage, radiotherapy, and chemotherapy) showed robust but limited discrimination in the validation cohort (AUC of approximately 0.50–0.75).²⁵ In addition, several machine learning-related studies, such as models based on inflammatory markers, although they performed well in prediction (AUC 0.95–0.92), relied on laboratory detection indicators, such as C-reactive protein, which had high data acquisition costs and were not suitable for early screening at the population level.²⁶ The variables selected in this study, such as age and BMI, were all from routine health surveys, and no additional invasive tests are required, enhancing the strength of clinical promotion.²⁵

The analysis of variable importance further uncovers the critical risk factors, with the results indicating that obesity (BMI ≥ 30 kg/m²) is the most significant risk factor. Studies have shown that adipose tissue can secrete various inflammatory factors, and long-term inflammatory stimulation may increase the risk of lymphoma carcinogenesis. Obesity can induce immune system dysfunction, leading to alterations in the metabolism and function of immune cells, which in turn impair their ability to recognize and eliminate cancer cells.²⁷ More importantly, leptin, as a core adipokine that is elevated in obesity, has been shown to directly promote lymphoma cell

proliferation, inhibit apoptosis, and enhance its invasive ability by binding to leptin receptors on the surface of lymphoma cells and activating key signaling pathways, such as phosphoinositide 3-kinase–protein kinase B,²⁸ becoming a key molecular bridge connecting obesity and the occurrence and development of lymphoma. At the same time, obesity-induced insulin resistance can upregulate insulin-like growth factor signaling, further amplifying the pro-proliferative effect of tumor cells.²⁹

The BMI index, defined as weight in kilograms divided by the square of height in meters, is clinically used to classify weight status: overweight is defined as a BMI ≥ 25 kg/m². Between 1980 and 2013, the worldwide prevalence of overweight or obese adults increased by 28%.³⁰ Previous studies reported that obese/overweight patients with lymphoma had poorer overall survival due to disease relapse rather than therapy-related toxicity.^{31,32} A higher burden of cardiovascular disease and metabolic dysfunction that occurs in overweight and obese patients may independently be associated with poorer outcomes.³³ Obesity is associated with an increased risk of cardiovascular disease and malignancy,³⁴ particularly hematologic malignancies.^{35–39} Individuals living with obesity are 39% more likely to develop leukemia³⁹ and 20% more likely to develop NHL³⁵ relative to individuals with a normal BMI (18.5 to ≥ 25 kg/m²). Several studies have shown an increased risk of NHL in individuals with

higher BMI, and the risk appears particularly notable for aggressive NHL subtypes, rather than indolent ones.^{35,36,38} Indeed, obesity has been associated with a 40% increased risk of developing DLBCL, whereas there is no statistically significant increase in the risk of developing follicular lymphoma or small lymphocytic lymphoma.³⁵

Our study shows that the risk of lymphoma increases with age, especially in people over 65 years old, who are particularly prone to lymphoma. The core biological mechanism underlying this phenomenon is closely related to age-related immunosenescence and genomic instability. With age, thymic degeneration leads to decreased T-cell production, an imbalance in lymphocyte subsets, a decline in B-cell function, a decrease in antibody diversity, and chronic inflammation. This dual change, weakened immune surveillance and chronic inflammation, not only reduces the body's ability to recognize and clear tumor cells, but also continuously stimulates the proliferation of lymphoma cells through inflammatory factors, such as interleukin-6 and tumor necrosis factor.^{40,41} In addition, age-related telomere shortening leads to increased genomic instability.⁴² Shortening of telomeres in peripheral blood leukocytes has been confirmed to be significantly associated with an increased risk of Hodgkin's lymphoma. The mechanism involves disruption of chromosome stability, a decline in DNA damage repair capacity, and the acceleration of malignant transformation of lymphocytes.⁴³

In summary, obesity (BMI ≥ 30 kg/m²) and age groups of 45–64 years and ≥ 65 years are significant risk factors for lymphoma. Obesity promotes the development of lymphoma by inducing chronic inflammation and immune dysfunction, particularly increasing the risk of aggressive subtypes (e.g., DLBCL), and obese patients have poorer overall survival.

Using lymphoma-related data from the NHIS database, this study identified statistically significant differences between lymphoma patients and controls by constructing a baseline characteristics table. We then developed four machine learning prediction models based on these variables and found that the XGBoost model performed best. The importance of the seven variables was assessed using the machine learning algorithm, and it was found that obesity (BMI ≥ 30 kg/m²), age 45–64 years, and age ≥ 65 were significant predictors of lymphoma. The subsequent variable importance ranking provides a reference value for the diagnosis and management of lymphoma. However, this study has several limitations, including its cross-sectional design, which captures information at a single point in time, making it difficult to establish causality or track changes over time. In addition, as a retrospective observational study, data collection relied on self-reported

or historical records, which may be incomplete and prone to biases.

5. Conclusion

Based on NHIS data, this study found that obesity made the greatest predictive contribution to lymphoma development, and the established prediction model demonstrated good discriminative performance. This research provided a basis for risk assessment, early prevention, and treatment of lymphoma.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Yuying Guo, Xuejing Yang, Dong Song

Investigation: Yuying Guo, Shuling Hou, Dong Song

Methodology: Yuying Guo, Xuejing Yang, Dong Song

Writing—original draft: Yuying Guo, Dong Song

Writing—review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data used in this study are publicly available on the official website of the National Health Interview Survey (NHIS) at <https://www.cdc.gov/nchs/nhis/index.html>.

References

1. Yu T, Xu-Monette ZY, Yu L, Li Y, Young KH. Mechanisms of ferroptosis and targeted therapeutic approaches in lymphoma. *Cell Death Dis.* 2023;14(11):771. doi: 10.1038/s41419-023-06295-w
2. Matasar MJ, Zelenetz AD. Overview of lymphoma diagnosis and management. *Radiol Clin North Am.* 2008;46(2):175–198, vii. doi: 10.1016/j.rcl.2008.03.005
3. Mugnaini EN, Ghosh N. Lymphoma. *Prim Care.* 2016;43(4):661–675.

- doi: 10.1016/j.pop.2016.07.012
4. China Anti-Cancer Association Lymphoma Committee, Chinese Association for Clinical Oncologists, Medical Oncology Branch of Chinese International Exchange and Promotion Association for Medical and Healthcare. Clinical practice guideline for multi-disciplinary treatment strategy of lymphoma in China. *Zhonghua Zhong Liu Za Zhi*. 2021;43(2):163-166.
doi: 10.3760/cma.j.cn112152-20201109-00971
 5. Melani C, Wilson WH. Front-line treatment of diffuse large B-Cell lymphoma in patients with cardiovascular comorbidities; Omission of anthracycline reduces cure. *Leuk Lymphoma*. 2022;63(3):511-513.
doi: 10.1080/10428194.2021.2002323
 6. Bhatt VR, Vose JM. Hematopoietic stem cell transplantation for non-Hodgkin lymphoma. *Hematol Oncol Clin North Am*. 2014;28(6):1073-1095.
doi: 10.1016/j.hoc.2014.08.015
 7. Mbous YPV, Mohamed R, Bhandari R. A decomposition analysis of racial disparities in physical activity among cancer survivors: National health interview survey 2009-2018. *J Phys Act Health*. 2023;20(8):760-771.
doi: 10.1123/jpah.2022-0356
 8. Moron LP, Irimata KE, Parker JD. Comparison of mental health estimates by sociodemographic characteristics in the research and development survey 3 and the 2019 national health interview survey. *Natl Health Stat Report*. 2023;188:1-11.
 9. Panos A, Mavridis D. TableOne: An online web application and R package for summarising and visualising data. *Evid Based Ment Health*. 2020;23(3):127-130.
doi: 10.1136/ebmental-2020-300162
 10. Wang Z, Liu Q, Nemes J, *et al*. Associations between blast exposures and intestinal permeability and neurotrauma symptoms during mortar fire military tactical training operations. *Mil Med*. 2025:usaf478.
doi: 10.1093/milmed/usaf478
 11. Yamamoto Y, Tsuzuki T, Akatsuka J, *et al*. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nat Commun*. 2019;10(1):5642.
doi: 10.1038/s41467-019-13647-8
 12. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevskaya O, Written on Behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med*. 2019;7(7):152.
doi: 10.21037/atm.2019.03.29
 13. Holland CH, Tanevski J, Perales-Patón J, *et al*. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol*. 2020;21(1):36.
doi: 10.1186/s13059-020-1949-z
 14. Collet E, Dalac S, Caillot D, Chavanet P, Beer F, Lambert D. [Paraneoplastic vasculitis and multiple myeloma]. *Presse Med*. 1990;19(23):1105.
 15. Fallon T, Nolan R, Peters J, Heron N. "Beyond the finish line" the epidemiology of injury and illness in professional cycling: Insights from a year-long prospective study. *Sports (Basel)*. 2025;13(1):20.
doi: 10.3390/sports13010020
 16. Didion JP, Morgan AP, Clayshulte AM, *et al*. A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLoS Genet*. 2015;11(2):e1004850.
doi: 10.1371/journal.pgen.1004850
 17. Lin G, Gao Z, Wu S, *et al*. scRNA-seq revealed high stemness epithelial malignant cell clusters and prognostic models of lung adenocarcinoma. *Sci Rep*. 2024;14(1):3709.
doi: 10.1038/s41598-024-54135-4
 18. Han S, Liu Y, Li X, *et al*. Development and validation of a preoperative nomogram for predicting benign and malignant gallbladder polypoid lesions. *Front Oncol*. 2022;12:800449.
doi: 10.3389/fonc.2022.800449
 19. Bakhshi TJ, Georgel PT. Genetic and epigenetic determinants of diffuse large B-cell lymphoma. *Blood Cancer J*. 2020;10(12):123.
doi: 10.1038/s41408-020-00389-w
 20. Zhuang S, Yang Z, Cui Z, Zhang Y, Che F. Epigenetic alterations and advancement of lymphoma treatment. *Ann Hematol*. 2024;103(5):1435-1454.
doi: 10.1007/s00277-023-05395-z
 21. Kendel NE, Stanek JR, Willen FK, Audino AN. Characterizing age-related differences in Hodgkin lymphoma in children, adolescents and young adults. *Pediatr Hematol Oncol*. 2024;41(5):336-345.
doi: 10.1080/08880018.2024.2337627
 22. Huang J, Pang WS, Lok V, *et al*. Incidence, mortality, risk factors, and trends for Hodgkin lymphoma: A global data analysis. *J Hematol Oncol*. 2022;15(1):57.
doi: 10.1186/s13045-022-01281-9
 23. Chang PW, Newman TB. Receiver operating characteristic (ROC) curves: The basics and beyond. *Hosp Pediatr*. 2024;14(7):e330-e334.
doi: 10.1542/hpeds.2023-007462
 24. Dong J, Peng L, Yang X, Zhang Z, Zhang P. XGBoost-based intelligence yield prediction and reaction factors analysis of amination reaction. *J Comput Chem*. 2022;43(4):289-302.
doi: 10.1002/jcc.26791

25. Wang J, Ren W, Zhang C, Wang X. A new staging system based on the dynamic prognostic nomogram for elderly patients with primary gastrointestinal diffuse large B-Cell lymphoma. *Front Med (Lausanne)*. 2022;9:860993.
doi: 10.3389/fmed.2022.860993
26. Li Y, Zhu Y, Duan X. Prognostic value of multiple immune inflammatory markers in diffuse large B-cell lymphoma. *Am J Transl Res*. 2023;15(4):2610-2621.
27. Vaitkus JA, Celi FS. The role of adipose tissue in cancer-associated cachexia. *Exp Biol Med (Maywood)*. 2017;242(5):473-481.
doi: 10.1177/1535370216683282
28. Uddin S, Bu R, Ahmed M, et al. Leptin receptor expression and its association with PI3K/AKT signaling pathway in diffuse large B-cell lymphoma. *Leuk Lymphoma*. 2010;51(7):1305-1314.
doi: 10.3109/10428191003802365
29. Arcidiacono B, Iiritano S, Nocera A, et al. Insulin resistance and cancer risk: An overview of the pathogenetic mechanisms. *Exp Diabetes Res*. 2012;2012:789174.
doi: 10.1155/2012/789174
30. Ng M, Fleming T, Robinson M, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: A systematic analysis for the global burden of disease study 2013. *Lancet*. 2014;384(9945):766-781.
doi: 10.1016/s0140-6736(14)60460-8
31. Breccia M, Mazzarella L, Bagnardi V, et al. Increased BMI correlates with higher risk of disease relapse and differentiation syndrome in patients with acute promyelocytic leukemia treated with the AIDA protocols. *Blood*. 2012;119(1):49-54.
doi: 10.1182/blood-2011-07-369595
32. Obesity: Preventing and managing the global epidemic. Report of a WHO consultation. *World Health Organ Tech Rep Ser*. 2000;894:i-xii, 1-253.
33. Mapp S, Sandhu G, Carrington C, Hennig S. A systematic review of treatment outcomes with weight-based dosing of chemotherapy in obese adult patients with acute leukemia or lymphoma. *Leuk Lymphoma*. 2016;57(4):981-984.
doi: 10.3109/10428194.2015.1087520
34. De Pergola G, Silvestris F. Obesity as a major risk factor for cancer. *J Obes*. 2013;2013:291546.
doi: 10.1155/2013/291546
35. Larsson SC, Wolk A. Obesity and risk of non-Hodgkin's lymphoma: A meta-analysis. *Int J Cancer*. 2007;121(7):1564-1570.
doi: 10.1002/ijc.22762
36. Larsson SC, Wolk A. Body mass index and risk of non-Hodgkin's and Hodgkin's lymphoma: A meta-analysis of prospective studies. *Eur J Cancer*. 2011;47(16):2422-2430.
doi: 10.1016/j.ejca.2011.06.029
37. Wallin A, Larsson SC. Body mass index and risk of multiple myeloma: A meta-analysis of prospective studies. *Eur J Cancer*. 2011;47(11):1606-1615.
doi: 10.1016/j.ejca.2011.01.020
38. Patel AV, Diver WR, Teras LR, Birmann BM, Gapstur SM. Body mass index, height and risk of lymphoid neoplasms in a large United States cohort. *Leuk Lymphoma*. 2013;54(6):1221-1227.
doi: 10.3109/10428194.2012.742523
39. Larsson SC, Wolk A. Overweight and obesity and incidence of leukemia: A meta-analysis of cohort studies. *Int J Cancer*. 2008;122(6):1418-1421.
doi: 10.1002/ijc.23176
40. Barbé-Tuana F, Funchal G, Schmitz CRR, Maurmann RM, Bauer ME. The interplay between immunosenescence and age-related diseases. *Semin Immunopathol*. 2020;42(5):545-557.
doi: 10.1007/s00281-020-00806-z
41. Thomas R, Wang W, Su DM. Contributions of age-related thymic involution to immunosenescence and inflammaging. *Immun Ageing*. 2020;17:2.
doi: 10.1186/s12979-020-0173-8
42. Burren OS, Dhindsa RS, Deevi SVV, et al. Genetic architecture of telomere length in 462,666 UK Biobank whole-genome sequences. *Nat Genet*. 2024;56(9):1832-1840.
doi: 10.1038/s41588-024-01884-7
43. M'Kacher R, Bennaceur-Griscelli A, Girinsky T, et al. Telomere shortening and associated chromosomal instability in peripheral blood lymphocytes of patients with Hodgkin's lymphoma prior to any treatment are predictive of second cancers. *Int J Radiat Oncol Biol Phys*. 2007;68(2):465-471.
doi: 10.1016/j.ijrobp.2007.01.050