Machine Learning-Based Survival Prediction Models for Progression-Free and Overall Survival in Advanced-Stage Hodgkin Lymphoma

Rask Kragh Jørgensen, Rasmus; Bergström, Fanny; Eloranta, Sandra; Tang Severinsen, Marianne; Bjøro Smeland, Knut; Fosså, Alexander; Christensen, Jacob Haaber; Hutchings, Martin; Bo Dahl-Sørensen, Rasmus; Kamper, Peter; Glimelius, Ingrid; Smedby, Karin E.; K Parsons, Susan.; Mae Rodday, Angie; J Maurer, Matthew.; M Evens, Andrew.; El-Galaly, Tarec C.; Hjort Jakobsen, Lasse

Published in:
JCO Clinical Cancer Informatics

DOI (link to publication from Publisher):
10.1200/CCI.23.00255

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Machine Learning–Based Survival Prediction Models for Progression-Free and Overall Survival in Advanced-Stage Hodgkin Lymphoma

Rasmus Rask Kragh Jørgensen, MSc[1,2]; Fanny Bergström, MSc[3]; Sandra Eloranta, PhD[3]; Marianne Tang Severinsen, MD, PhD[1,2]; Knut Bjøro Smeland, MD, PhD[4]; Alexander Fosså, MD, PhD[4]; Jacob Haaber Christensen, MD, PhD[5]; Martin Hutchings, MD, PhD[6,7]; Rasmus Bo Dahl-Sørensen, MD[8]; Peter Kamper, MD, PhD[9]; Ingrid Glimelius, MD, PhD[3,10]; Karin E. Smedby, MD, PhD[3,11]; Susan K. Parsons, MD, MRP[12]; Angie Mae Rodday, MSc, PhD[12]; Matthew J. Maurer, MSc, DMSc[13]; Andrew M. Evens, DO, MBA, MSc[14]; Tarec C. El-Galaly, MD, DMSc[1,2]; and Lasse Hjort Jakobsen, MSc, PhD[1,15]

DOI https://doi.org/10.1200/CCI.23.00255

ABSTRACT

PURPOSE Patients diagnosed with advanced-stage Hodgkin lymphoma (aHL) have historically been risk-stratified using the International Prognostic Score (IPS). This study investigated if a machine learning (ML) approach could outperform existing models when it comes to predicting overall survival (OS) and progression-free survival (PFS).

PATIENTS AND METHODS This study used patient data from the Danish National Lymphoma Register for model development (development cohort). The ML model was developed using stacking, which combines several predictive survival models (Cox proportional hazard, flexible parametric model, IPS, principal component, penalized regression) into a single model, and was compared with two versions of IPS (IPS-3 and IPS-7) and the newly developed aHL international prognostic index (A-HIPI). Internal model validation was performed using nested cross-validation, and external validation was performed using patient data from the Swedish Lymphoma Register and Cancer Registry of Norway (validation cohort).

RESULTS In total, 707 and 760 patients with aHL were included in the development and validation cohorts, respectively. Examining model performance for OS in the development cohort, the concordance index (C-index) for the ML model, IPS-7, IPS-3, and A-HIPI was found to be 0.789, 0.608, 0.650, and 0.768, respectively. The corresponding estimates in the validation cohort were 0.749, 0.700, 0.663, and 0.741. For PFS, the ML model achieved the highest C-index in both cohorts (0.665 in the development cohort and 0.691 in the validation cohort). The time-varying AUCs for both the ML model and the A-HIPI were consistently higher in both cohorts compared with the IPS models within the first 5 years after diagnosis.

CONCLUSION The new prognostic model for aHL on the basis of ML techniques demonstrated a substantial improvement compared with the IPS models, but yielded a limited improvement in predictive performance compared with the A-HIPI.

ACCOMPANYING CONTENT

Data Supplement

Accepted February 21, 2024
Published April 12, 2024

JCO Clin Cancer Inform
8:e2300255

INTRODUCTION

Hodgkin lymphoma (HL) is among the most common malignancies in younger individuals and is historically divided into classical HL (cHL; approximately 95%) and nodular lymphocyte–predominant HL (approximately 5%).[1,2] In cHL, cure rates and outcomes across disease stages are generally excellent because of combined modality treatment for limited-stage disease and effective multiagent chemotherapy regimens for advanced-stage disease.[3] Despite these advancements, outcomes still remain variable and 10%-20% of patients will relapse or have refractory disease.[1] A risk-adapted treatment strategy requires accurate prognostic scores on the basis of factors that are available at diagnosis to select the best treatment for an individual patient with HL. With accurate prognostic scores, clinicians can reserve the most intensive and toxic treatment for high-risk patients who need it and spare low-risk patients from unnecessary toxicity. Hence, precise and updated prognostic scores are important for optimal treatment selection for an individual

CONTEXT

Key Objective

To build machine learning (ML) models that predict overall survival and progression-free survival in patients with advanced-stage Hodgkin lymphoma (aHL) and validate it against the International Prognostic Score (IPS) and the newly developed aHL international prognostic index (A-HIPI).

Knowledge Generated

The ML models were able to outperform the IPS in both internal and external validation but yielded limited improvement in predictive performance compared with the A-HIPI. The ML model and the A-HIPI were able to identify high-risk patients with worse survival than those defined as high-risk by the IPS.

Relevance *(J.L. Warner)*

ML-based prognostic algorithms could be implemented at the point-of-care. While in this case the ML model only slightly outperformed the newly developed A-HIPI, it is nevertheless a proof-of-principle for the approach.*

*Relevance section written by *JCO* Editor-in-Chief Jeremy L. Warner, MD, MS, FAMIA, FASCO.

patient. The International Prognostic Score (IPS-7) is a commonly used risk-stratification score for advanced-stage HL (aHL).[4,5] The IPS-7 consists of seven clinical risk factors and identifies six risk groups. The IPS-3 model (including only age, Ann Arbor stage, and hemoglobin level) was introduced to simplify the IPS-7 while preserving predictive accuracy.[5,6] To simplify the use of IPS-7 and IPS-3, these have been constructed by dichotomizing important prognostic variables and thus provide a risk classification for individual patients. However, the dichotomization of the underlying variables used in both IPS-7 and IPS-3 leads to substantial loss of information, which can result in reduced accuracy of the provided prognostic information.[7] The aHL international prognostic index (A-HIPI)[8] was recently developed using Cox proportional hazards (CPH) models without dichotomizing continuous or categorical clinical factors and indeed demonstrated improvements in predictive capability compared with the IPS models. In the current study, new prognostic models for aHL were developed using machine learning (ML) and an extended list of baseline clinical factors with the aim of further improving predictive performance, which has shown great utility in previous studies for other diseases.[9,10]

PATIENTS AND METHODS

Patients

The ML models were developed using clinical data from the nationwide Danish National Lymphoma Register (LYFO). LYFO contains information on >25,000 patients with lymphoma, including subtype, diagnostic characteristics, laboratory values, treatment, and outcomes.[11,12] Adult patients (18 years and older at diagnosis) who fulfilled the following criteria were included in the study: (1) newly diagnosed cHL, (2) advanced-stage disease defined by Ann Arbor stage III-IV or IIB with extranodal disease,[1] and (3) receipt of curative-intent treatment with BEACOPP (bleomycin, etoposide, doxorubicin hydrochloride, cyclophosphamide, vincristine, procarbazine, and prednisone) or ABVD (doxorubicin, bleomycin, vinblastine, and dacarbazine).

Patients diagnosed between 2006 and 2020 were included from LYFO (referred to as the development cohort). The model was validated using data from the Cancer Registry of Norway and the Swedish Lymphoma Register[13] (SLR). Clinical information for Norwegian patients was extracted from the clinical lymphoma database at Oslo University Hospital (CLDOUH). Validation studies have shown high coverage of the Swedish SLR, the cancer registry of Norway, and Danish LYFO registers.[11,14,15] Patients recorded in the Swedish and Norwegian registers from 2007 until 2020 and 2017, respectively, who met the abovementioned inclusion criteria were pooled to form the validation cohort. Patients with missing information on >2 of the 22 clinical factors included in the models were excluded from the study.

Model Development

Overall survival (OS) was defined as the time from HL diagnosis until death or censoring, and progression-free survival (PFS) was defined as the time from HL diagnosis until progression, relapse, death, or censoring. Censoring occurred on the last date of follow-up or December 31, 2020, in the Danish cohort, December 31, 2020, in the Swedish cohort, and December 31, 2017, in the Norwegian cohort. For the validation cohort, exact relapse and progression dates were not complete for all patients. For these, progression dates were imputed using the last date of first-line treatment and relapse dates were imputed using the start date of second-line treatment.

The ML models for predicting OS and PFS were built using an ensemble method called stacking.[16,17] The stacking method combines a number of survival models into a single survival model through a weighted average approach:

$$S_{\text{Stack}}(t) = \sum_{j=1}^{m} \alpha_j S_j(t), \quad \text{where } \sum_{j=1}^{m} \alpha_j = 1 \text{ and } \alpha_j \in [0,1]$$

where $S_j(t)$ represents the survival probability for the jth model and m is the number of models included. The stacking weights, α_j, were determined by minimizing the integrated brier score (IBS)[18] obtained through cross-validation (CV; Data Supplement, Fig S1).

The following survival models were included in the stacking procedure: a CPH model, penalized CPH model,[19] principal component[20] CPH model, IPS-7, IPS-3, and flexible parametric survival (FPS) models.[21] The models included an extended list of clinical factors, depending on the model type (Data Supplement, Table S1). The survival models incorporated in the ML models, their input, and model abbreviation are described in Table 1. The models were selected to include a broad range of different survival models, including models with time-varying covariate effects.[22]

The initial covariate selection was done by aligning data availability in all cohorts, and only covariates present in all cohorts were included (Data Supplement, Section A4). Missing data on predictors were handled using single value imputation before model development. Further details about the imputation process and methods used for this purpose can be found in the Data Supplement (Section A5).

Model Performance and Validation

Performance measures used were the IBS, time-varying AUC,[23,24] and the concordance index (C-index).[25,26] These measures were evaluated for the first 5 years after diagnosis. Performance measures for the development cohort were calculated by pooling predicted probability estimates in a 10-fold CV setup (Data Supplement, Fig S1). A nested CV setup was applied to avoid overly optimistic performance measures.[27]

To evaluate the ability of the ML models and A-HIPI to identify patients with poor outcomes, patients were ranked on the basis of their predicted 5-year OS and PFS. The IPS high-risk group (IPS score of 5-7) was then compared with the patients with the lowest predicted probability at 5 years (ML high-risk, and A-HIPI high-risk). The probability threshold for the new high-risk group was defined such that the ML high-risk group and the IPS high-risk group included the same number of patients. Outcomes of the two high-risk groups were compared visually using the Kaplan-Meier method. A similar analysis was performed for patients at low risk.

Model calibration was assessed using calibration plots, calibration slope (CS), and calibration-in-the-large (CITL)[28] and was performed in the validation cohort at 5 years postdiagnosis. To investigate the validation of the A-HIPI further, a separate calibration analysis was performed for young patients (age 18-65 years). All analyses were performed using R version 4.1.1 (R Foundation for Statistical Computing, Vienna, Austria; Data Supplement, Table S2).

The study was approved by the North Denmark Region (ID: 2021-002), the Swedish Regional ethics committee in Stockholm (Dnr 2019-00242), and the Regional Committee for Medical Research Ethics South East Norway (No. 2018/2209).

RESULTS

Among a total of 4,582 patients with cHL, 3,115 were excluded (Fig 1) and 1,467 patients were included (707 in the

TABLE 1. The Prediction Models Incorporated in the ML Model

Model	Type	Variable Subset	Notes
CPH-Enet	CPH model with elastic net regularization	1	CV was applied to estimate regularization parameters
CPH-LASSO	CPH model with LASSO regularization	1	Same as CPH-Enet
CPH-LASSO-Int	CPH-LASSO including all possible first-order interaction terms	1	Same as CPH-Enet
PC-CPH	CPH model with PC vectors as predictors	1	A PC rotation was applied to all covariates
FPS	FPS models with time-varying covariates	2	Coefficient estimates for the baseline and time-varying coefficients were modeled by natural cubic splines with 2 degrees of freedom
CPH-IPS	CPH model with IPS covariates	3	IPS covariates were used in their continuous form
IPS-7	IPS risk group Kaplan-Meier estimates	3	
IPS-3	IPS risk group Kaplan-Meier estimates with 3 covariates	4	
FPS-age-PS	FPS model with age and PS as time-varying covariates	5	Same as for the FPS model

Abbreviations: CPH, Cox proportional hazard; CV, cross-validation; Enet, elastic net; FPS, flexible parametric survival; Int, interactions; IPS, International Prognostic Score; LASSO, least absolute shrinkage and selection operator; ML, machine learning; PC, principal component; PS, performance status.

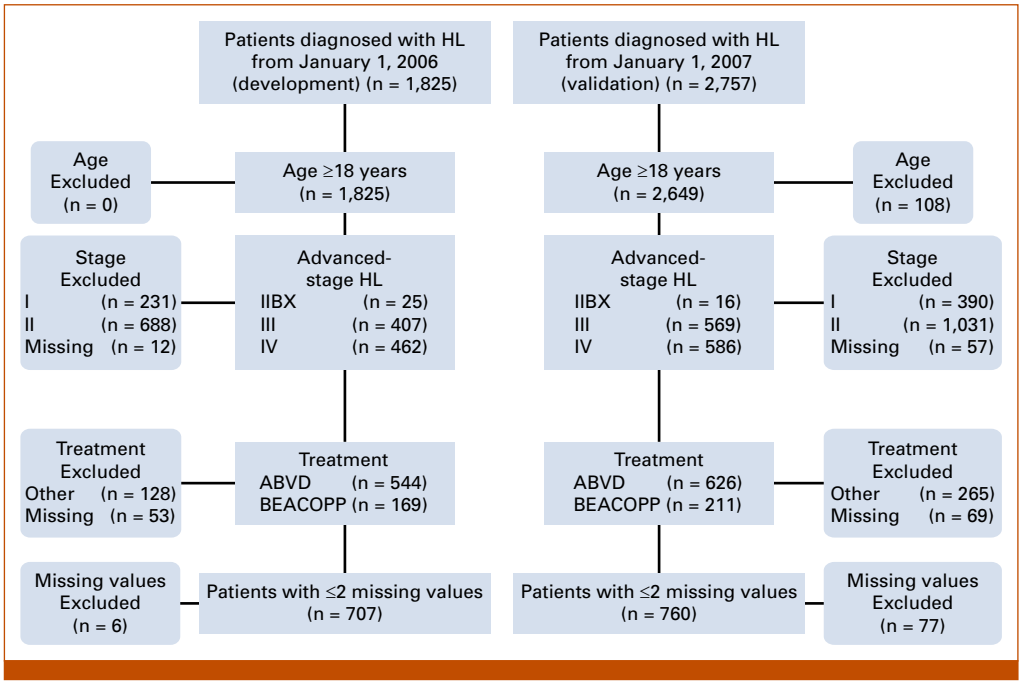

FIG 1. CONSORT diagram of patient flow in the development and validation cohorts. ABVD, doxorubicin, bleomycin, vinblastine, and dacarbazine; BEACOPP, bleomycin, etoposide, doxorubicin hydrochloride, cyclophosphamide, vincristine, procarbazine, and prednisone; HL, Hodgkin lymphoma; IIBX, Ann Arbor stage II with B-symptoms and extranodal involvement.

development cohort and 760 in the validation cohort). The median follow-up (reverse Kaplan-Meier method[29]) was 7.2 years in the development cohort. The 5-year OS was 83% (95% CI, 81 to 87), and the 5-year PFS was 73% (95% CI, 70 to 77; Data Supplement, Fig S2A). For the validation cohort, the median follow-up was 8.6 years. The 5-year OS was 91% (95% CI, 88 to 93), and the PFS was 87% (95% CI, 85 to 90; Data Supplement, Fig S2B). Clinical characteristics in the cohorts were similar except for age (Table 2 and Data Supplement, Table S3). The weights in the ML models showed that a small number of models obtained non-negligible weights (α_j > 0.001, Data Supplement, Table S4).

Validation

IPS Versus ML

For OS in the development cohort, the ML model, IPS-7, and IPS-3 achieved a C-index of 0.789, 0.608, and 0.650, respectively. In the validation cohort, the corresponding C-index was 0.749, 0.700, and 0.663, respectively. For PFS, the C-index in the development cohort was 0.665 for the ML model, 0.549 for the IPS-7, and 0.576 for the IPS-3. The corresponding estimates in the validation cohort were 0.691 for the ML model, 0.672 for the IPS-7, and 0.640 for the IPS-3. The ML model attained a lower IBS for OS in both the development and validation cohorts compared with the IPS-7 and IPS-3 except for PFS in the validation cohort (Table 3). For both OS and PFS, the time-varying AUC in the development cohort was consistently higher for the ML model compared with the IPS-7 and IPS-3 (Fig 2A). Similar results were seen in the validation cohort (Fig 2B).

Patients classified as high risk (13.9% of patients) in the development cohort, according to the ML model, had a 5-year OS of 51% (95% CI, 41 to 63), whereas the 5-year OS was 70% (95% CI, 61 to 80) in patients defined as high risk by the IPS-7. The 5-year PFS was 50% (95% CI, 40 to 62) for the ML high-risk group and 58% (95% CI, 49 to 69) for the IPS-7 high-risk group (Fig 3A). In the validation cohort, patients in the ML high-risk group (9.2% of patients) had a 5-year OS of 70% (95% CI, 59 to 81), whereas patients defined as high risk by the IPS-7 had a 5-year OS of 74% (95% CI, 64 to 86). The corresponding estimates for PFS were 65% (95% CI, 55 to 78) and 69% (95% CI, 60 to 82), respectively (Fig 3B). Patients classified as low risk in both cohorts, according to the ML model, had similar OS and PFS as the IPS-7 (Data Supplement, Figs S3A and S3B).

A-HIPI Versus ML

In the development cohort, the A-HIPI achieved a C-index of 0.768 for OS and 0.649 for PFS. The corresponding estimates for the validation cohort were 0.741 for OS and 0.677 for PFS. The time-varying AUC was similar to the ML model for OS and PFS in the development cohort (Fig 2A). For the validation cohort, the time-varying AUC estimates for A-HIPI were similar for OS but attained similar estimates as the IPS-7 for PFS (Fig 2B). The CITL and CS for the A-HIPI in the

TABLE 2. Characteristics of Patients With Advanced Hodgkin Lymphoma in the Development and Validation Cohort

Clinical Factor	Development Cohort (n = 707)	Validation Cohort (n = 760)
Age, median (range)	44 (18-87)	37 (18-74)
Sex, No. (%)		
Female	292 (41.3)	295 (38.8)
Male	415 (58.7)	465 (61.2)
Ann Arbor, No. (%)		
IIB	20 (2.8)	10 (1.3)
III	324 (45.8)	360 (47.4)
IV	363 (51.3)	390 (51.3)
PS, No. (%)		
0 and missing[a]	466 (65.9)	484 (63.7)
1	187 (26.4)	217 (28.6)
2	27 (3.8)	42 (5.5)
3	18 (2.5)	14 (1.8)
4	9 (1.3)	3 (0.4)
B-symptoms, No. (%)		
No	234 (33.1)	291 (38.3)
Yes	462 (65.3)	460 (60.5)
Missing	11 (1.6)	9 (1.2)
Nodal, median (range)	4 (0-11)	4 (0-15)
Extranodal, median (range)	1 (0-5)	0 (0-4)
Tumor diameter, No. (%)		
<10 cm	505 (71.4)	564 (74.2)
≥10 cm	87 (12.3)	124 (16.3)
Missing	115 (16.3)	72 (9.5)
Treatment, No. (%)		
ABVD	538 (76.1)	566 (74.5)
BEACOPP	169 (23.9)	194 (25.5)
Radiotherapy, No. (%)		
No	621 (87.8)	676 (88.9)
Yes	86 (12.2)	84 (11.1)

Abbreviations: ABVD, doxorubicin, bleomycin, vinblastine, and dacarbazine; BEACOPP, bleomycin, etoposide, doxorubicin hydrochloride, cyclophosphamide, vincristine, procarbazine, and prednisone; PS, WHO performance status.
[a]Missing values are pooled together with PS 0 because of data sharing regulations. In total, there were 21 missing values for PS.

TABLE 3. Performance Measures for the ML Model, IPS-3, IPS-7, and A-HIPI in the Development and Validation Cohorts

	Development Cohort				Validation Cohort			
	OS		PFS		OS		PFS	
Model	IBS	C-Index	IBS	C-Index	IBS	C-Index	IBS	C-Index
ML	0.073	0.789	0.141	0.665	0.049	0.749	0.085	0.691
IPS-3	0.085	0.650	0.149	0.576	0.054	0.663	0.078	0.64
IPS-7	0.087	0.608	0.151	0.549	0.052	0.700	0.077	0.672
A-HIPI	0.076	0.768	0.142	0.649	0.052	0.741	0.091	0.677

Abbreviations: A-HIPI, advanced-stage Hodgkin lymphoma international prognostic index; C-index, concordance index; IBS, integrated brier score; IPS, International Prognostic Score; ML, machine learning; OS, overall survival; PFS, progression-free survival.

development cohort were −0.14 and 0.99, respectively, for OS and 0.09 and 1.22, respectively, for PFS (Data Supplement, Fig S4). In the validation cohort, the ML model for OS attained CITL and CS values of 0.14 and 1.02, respectively, whereas the A-HIPI for OS attained CITL and CS values of 0.27 and 0.88, respectively. The ML model for PFS attained a CITL of 0.86 and a CS of 1.30, whereas the CITL and CS for the A-HIPI model for PFS were 0.89 and 1.61, respectively (Data Supplement, Figs S5 and S6). Restricting to patients age 18-65 years results in both cohorts remaining largely unchanged (Data Supplement, Figs S7 and S8). No substantial difference was found for patients classified as low-risk compared with the IPS-7 (Data Supplement, Figs S3A and S3B).

DISCUSSION

In the present study, we developed and validated a new prognostic model for patients with aHL using ML models that included more details on clinical factors than the conventional risk scores used for HL. The model outcompeted the conventional IPS-7 and IPS-3 in the development and validation cohorts with respect to both OS and PFS predictions. The ML model identified a high-risk group with a worse OS than the high-risk group identified through IPS-7. The improved performance of the ML model compared with IPS-7 and IPS-3 is not unexpected as the ML model uses a more flexible modeling approach and avoids dichotomization of clinical factors, thereby leading to more accurate prediction of outcomes. However, the ML models developed in this study exhibited only a minor improvement in performance compared with the A-HIPI.

Prognostic tools, such as the proposed model, have the potential to assist clinicians in patient counseling and the development of future trials for improved risk-adapted treatment approaches. The use of prognostic tools allows clinicians to identify individuals with low predicted survival probabilities who may not respond well to standard treatments and could benefit from novel treatment strategies. Conversely, for those with the best outcomes, trials may aim at treatment reduction to minimize the risk of long-term side effects. With the evolvement of new therapeutic options, precise risk is crucial, particularly for those with the highest risk. Thus, the introduction of new treatments that substantially change disease outcomes provides a need for updating and revalidating the existing prognostic models. This will also be the case for the ML model and the A-HIPI model as new treatments for aHL are introduced, particularly, as immune checkpoint inhibitors (anti-PD1) are made available in the first-line setting.

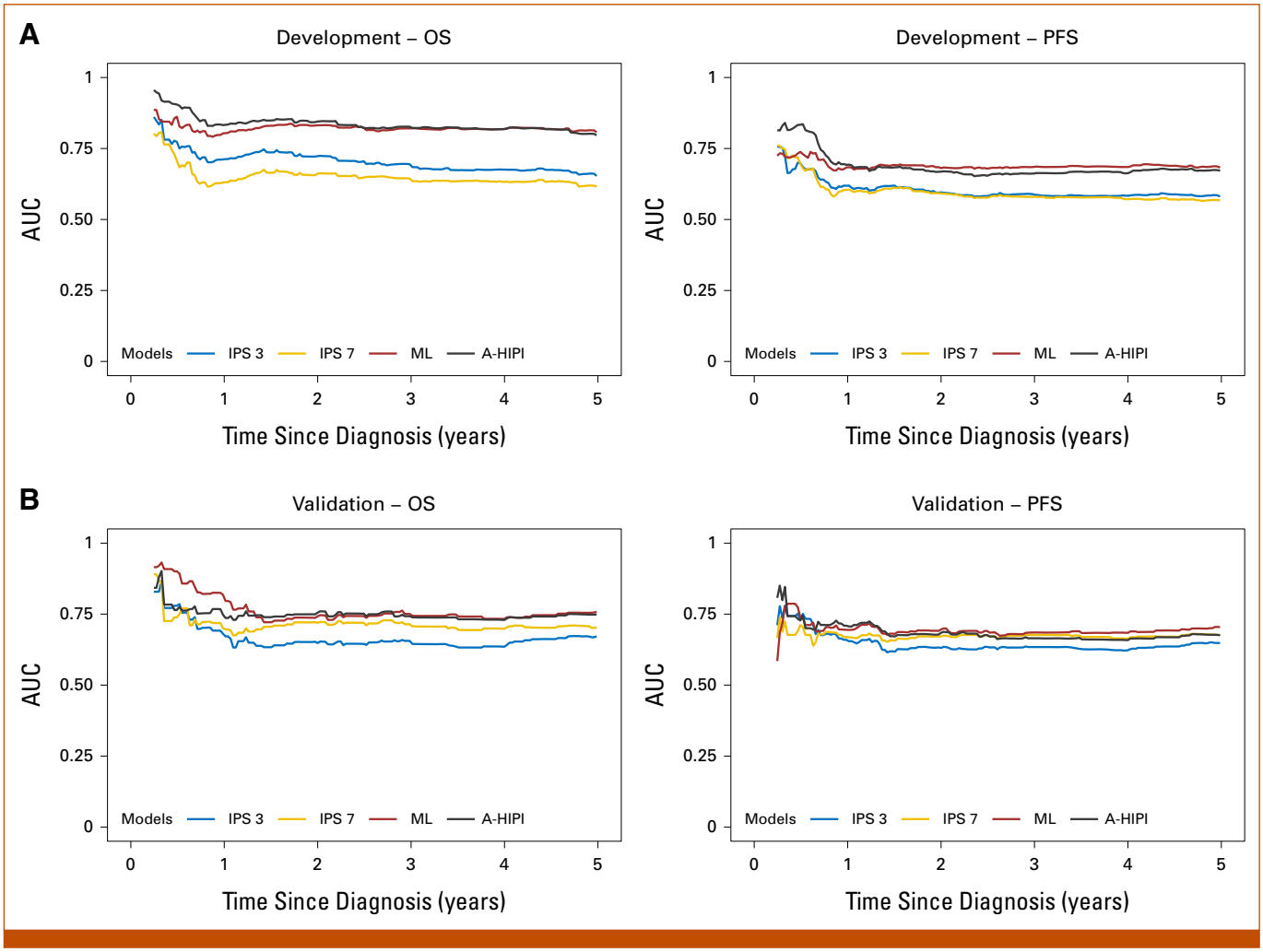

FIG 2. (A) Time-varying AUC for predicting OS and PFS for the ML model, the IPS-7, IPS-3, and A-HIPI. For the development cohort, the calculations for the ML model were performed by pooling predictions from each validation set in the cross-validation scheme. (B) Time-varying AUC for predicting OS and PFS for the ML model, the IPS-7, IPS-3, and A-HIPI in the validation cohort. A-HIPI, advanced-stage Hodgkin lymphoma international prognostic index; IPS, International Prognostic Score; ML, machine learning; OS, overall survival; PFS, progression-free survival.

Personalized treatment with response-adapted treatment strategies has been introduced to decrease long-term therapy-related toxicity and improve outcomes.[5,30-34] Recent trials in aHL have shown the challenge of identifying patients at sufficient risk and assessing the treatment effect in a group of patients with an overall favorable prognosis. For instance, both the HD15 and HD18 trials have required a high number of patients to formally show noninferiority of strategies with response-adapted reduction of treatment.[32,35] Similarly, the ECHELON-1 trial required over 1,300 patients to show a 6.9% increase in PFS with brentuximab vedotin, doxorubicin, vinblastine, and dacarbazine (A + AVD) over ABVD, a result that has not led to change of practice in all countries in the western world. The recent SWOG-S1826 trial included 994 patients and showed an approximately 8% increase in 1-year PFS in favor of nivolumab-AVD over A + AVD. From a statistical point of view, obtaining a sufficient number of events is crucial for obtaining robust estimates of the treatment effect in clinical trials, which is especially relevant within HL. Improving prognostic tools to more accurately identify patients at the highest risk could thus also lower the number of patients needed in aHL trials. The A-HIPI was recently developed as a new prognostic model for aHL in the modern era. The A-HIPI was developed using data from eight international phase III clinical trials and validated in four international cancer registries, demonstrating high performance compared with the IPS-7 and IPS-3. In our study, the A-HIPI outperformed the IPS models on most performance metrics for both OS and PFS, with the exception of a slightly higher IBS for PFS in the validation cohort (Table 3). This confirms the generalizability of the A-HIPI, which may be replacing the conventional IPS.

Calibration in the development cohort was better than that in the validation cohort for both the ML model and A-HIPI, especially for PFS where predictions from both models were substantially lower than the observed PFS in the validation

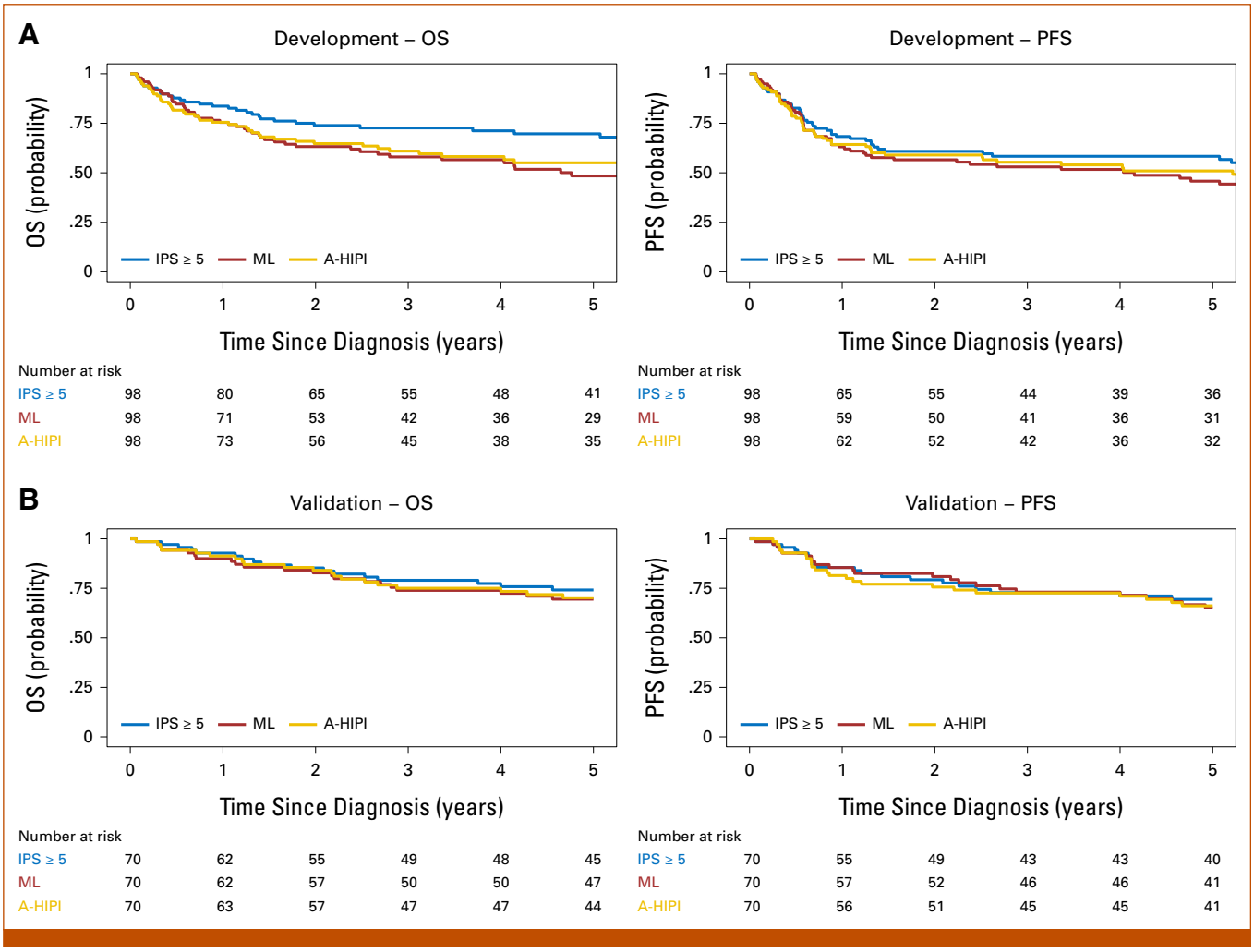

FIG 3. (A) Kaplan-Meier estimates of OS and PFS for high-risk patients (IPS-7, 5-7) and the same number of patients with the lowest predicted probability according to the ML model and the A-HIPI in the development cohort. (B) Kaplan-Meier estimates of OS and PFS for high-risk patients (IPS-7, 5-7) and the same number of patients with the lowest predicted probability according to the ML model in the validation cohort. A-HIPI, advanced-stage Hodgkin lymphoma international prognostic index; IPS, International Prognostic Score; ML, machine learning; OS, overall survival; PFS, progression-free survival.

cohort (Data Supplement, Figs S5 and S6). This may be explained by patient selection in the validation cohort, which is best exemplified by the higher OS and PFS in the validation cohort compared with the development cohort and the rather large difference of 7 years in median age at diagnosis. A substantial proportion of patients in the validation cohort had incomplete treatment information and were therefore excluded. This might have led to important differences in the development and validation cohorts that directly affect performance metrics. In addition, exact progression dates were incomplete in the validation cohort, and for patients with disease progression without an exact progression date, the progression date was imputed using the end of treatment date. The observed differences between the development and validation cohorts limit the generalizability of the external validation results seen. This data incompleteness limits analyses of external validation and highlights some of the challenges of using nationwide registries. In addition, the A-HIPI was only developed for patients age 18-65 years and is thus not designed for older patients, which might have also affected model performance in the present study. However, when limiting the cohort to 18- to 65-year-olds, results regarding calibration were largely unchanged (Data Supplement, Figs S7 and S8).

In this study, it was investigated whether a more flexible modeling strategy could increase the predictive performance over the IPS-7, IPS-3, and A-HIPI. Consistently, with a previous study within diffuse large B-cell lymphoma (DLBCL), we found that avoiding dichotomization and applying a flexible model strategy led to a substantial improvement in predictive performance.[9,36] However, the DLBCL study also found that compared with simple prognostic models for survival data, using ML to model outcomes only led to small increases in predictive performance. A similar finding was made in the present study where only limited differences between the developed ML model and the A-HIPI model were seen. Given the additional complexity of the ML approach developed in the current study, this suggests that ML models may hold limited

utility for outcome predictions in lymphoma when only a limited number of clinical factors are available to inform predictions. The clinical management of patients with aHL could potentially be guided by having accurate prognostic models for other survivorship end points. For instance, providing accurate prognostic information for patients in complete remission could inform the choice of post-treatment follow-up duration and intensity, whereas models for other important survivorship end points (like secondary malignancies and cardiovascular disease) would be helpful in planning individual patient follow-up programs. With the wide range of data that could be collected for aHL, ML models still have an important role to play in developing tools for these end points.

In the ML models, the most dominant underlying model (stacking weights; $\alpha_j = 0.807$; Data Supplement, Table S4) for OS was an FPS model including only age and WHO performance status. For PFS, the most important submodels (stacking weights; $\alpha_j > 0.1$) were the same FPS model, the CPH-LASSO-Int (least absolute shrinkage and selection operator with all possible interaction terms), and the CPH-IPS model, suggesting that accurate prediction of PFS requires more data points compared with OS. However, no model performed as well for PFS as they do for OS. This indicates that the clinical factors considered in the models can only to a limited degree predict response mechanisms in aHL or that patients with progressive aHL are kept alive for a long time because of effective salvage treatments. In addition, we found that using a CPH model including the continuous version of the IPS variables led to high performance metrics compared with all models in terms of both OS and PFS (Data Supplement, Table S5). This suggests that a limited number of clinical factors constitute the most important prognostic information for disease control in patients with aHL. Although this study focused on using clinical variables available at the time of diagnosis to predict the outcome of patients with aHL, studies in other lymphoma entities have shown that genetic information and molecular features are predictive. Thus, incorporating data such as cell of origin, genomics, radiomics, and histology data may improve performance beyond what can be achieved by the A-HIPI and similar models.[37-40] In addition, including positron emission tomography and computed tomography scan data, circulating tumor DNA would allow for a more refined tumor size staging system in regard to disease burden and has been shown to have great prognostic ability in HL.[41] These data were not available for the current study but may potentially improve PFS predictions in the future.

The strength of this study is the use of nationwide data with cohorts that are less prone to selection bias and is more representative of the entire HL patient population, ultimately improving measures of calibration. In addition, the few missing values in the data set were imputed on the basis of other clinical factors available at the time of diagnosis. Despite registry-based data from three countries, the limited number of patients with aHL and the high OS and PFS are not ideal for ML modeling and may thus affect the overall benefit of using more advanced predictive models. In addition, progression and relapse may not be fully captured in both the development and validation cohorts, but the excellent calibration measures of A-HIPI in the development cohort indicate that the missingness may be negligible for that cohort.

In conclusion, we have developed a new prognostic model using ML and compared this with existing prognostic models for aHL. The model provides personalized predictions on the basis of patient characteristics available at the time of diagnosis and is able to outperform the existing IPS models with respect to OS and PFS. However, the ML model only performs slightly better than the recently developed A-HIPI and a CPH model including the continuous version of the IPS variables. Given the simplicity of the A-HIPI, this model may be preferable to use in clinical practice.

AFFILIATIONS

[1]Department of Hematology, Clinical Cancer Research Centre, Aalborg University Hospital, Aalborg, Denmark
[2]Department of Clinical Medicine, Aalborg University, Aalborg, Denmark
[3]Clinical Epidemiology Division, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden
[4]Department of Oncology, Oslo University Hospital, Oslo, Norway
[5]Department of Hematology, Odense University Hospital, Odense, Denmark
[6]Department of Hematology, Rigshospitalet, Copenhagen, Denmark
[7]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
[8]Department of Hematology, Zealand University Hospital, Roskilde, Denmark
[9]Department of Hematology, Aarhus University Hospital, Aarhus, Denmark
[10]Department of Immunology, Genetics and Pathology, Cancer Precision Medicine, Uppsala University, Uppsala, Sweden
[11]Department of Hematology, Karolinska University Hospital, Stockholm, Sweden
[12]Department of Medicine, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA
[13]Department of Qualitative Health Sciences, Mayo Clinic, Rochester, MN
[14]Division of Blood Disorders, Rutgers Cancer Institute New Jersey, New Brunswick, NJ
[15]Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark

CORRESPONDING AUTHOR

Rasmus Rask Kragh Jørgensen, MSc; Twitter: @RasmusRaskKragh; e-mail: Rasmus.rask@rn.dk.

PRIOR PRESENTATION

Presented in part at the European Hematology Association 2022 congress, Vienna, Austria, June 9-17, 2022, and the American Society of Hematology 2023 Annual meeting, San Diego, CA, December 9-12, 2023.

SUPPORT

Supported by the Danish Cancer Society (R340-A19643 and R274-A17146-B1891) and the Nordic Cancer Union (R278-A15872 and R279-A15937).

DATA SHARING STATEMENT

Data are only accessible through the official data owners in Denmark, Sweden, and Norway. Therefore, it is not possible to share the data used in this study.

AUTHOR CONTRIBUTIONS

Conception and design: Rasmus Rask Kragh Jørgensen, Tarec C. El-Galaly, Lasse Hjort Jakobsen
Administrative support: Rasmus Rask Kragh Jørgensen, Marianne Tang Severinsen
Provision of study materials or patients: Jacob Haaber Christensen, Ingrid Glimelius, Karin E. Smedby, Susan K. Parsons
Collection and assembly of data: Rasmus Rask Kragh Jørgensen, Fanny Bergström, Marianne Tang Severinsen, Knut Bjøro Smeland, Alexander Fosså, Jacob Haaber Christensen, Martin Hutchings, Rasmus Bo Dahl-Sørensen, Ingrid Glimelius, Karin E. Smedby, Tarec C. El-Galaly, Lasse Hjort Jakobsen
Data analysis and interpretation: Rasmus Rask Kragh Jørgensen, Fanny Bergström, Sandra Eloranta, Knut Bjøro Smeland, Alexander Fosså, Peter Kamper, Karin E. Smedby, Susan K. Parsons, Angie Mae Rodday, Matthew J. Maurer, Andrew M. Evens, Tarec C. El-Galaly, Lasse Hjort Jakobsen
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Alexander Fosså
Honoraria: BMS Norway, Gilead Sciences, AbbVie, Takeda

Martin Hutchings
Consulting or Advisory Role: Takeda, Roche, Genmab, Janssen, AbbVie
Research Funding: Celgene (Inst), Genmab (Inst), Roche (Inst), Takeda (Inst), Novartis (Inst), Janssen (Inst), Merck (Inst), AbbVie (Inst), AstraZeneca (Inst)

Rasmus Bo Dahl-Sørensen
Stock and Other Ownership Interests: Bavarian Nordic
Travel, Accommodations, Expenses: Takeda

Peter Kamper
Honoraria: Pfizer
Travel, Accommodations, Expenses: Roche, Takeda

Ingrid Glimelius
Speakers' Bureau: Janssen-Cilag
Research Funding: Takeda (Inst)
Other Relationship: AbbVie (Inst)

Karin E. Smedby
Research Funding: Janssen-Cilag

Susan K. Parsons
Consulting or Advisory Role: Seagen

Matthew J. Maurer
Employment: Exact Sciences
Stock and Other Ownership Interests: Exact Sciences
Consulting or Advisory Role: AstraZeneca, BMS (Inst)
Research Funding: Bristol Myers Squibb (Inst), Roche/Genentech (Inst), Genmab (Inst)

Andrew M. Evens
Honoraria: Seagen, Pharmacyclics, Research to Practice, Epizyme, Novartis, MorphoSys, Curio Science, AbbVie/Pharmacyclics, Takeda, HUTCHMED, Incyte, Daiichi Sankyo/Astra Zeneca
Consulting or Advisory Role: Seagen, Novartis, Pharmacyclics, Miltenyi Biotec, Epizyme, MorphoSys, Cota Healthcare, AbbVie, Incyte
Speakers' Bureau: Research to Practice, Curio Science
Travel, Accommodations, Expenses: Seagen, Novartis, Curio Science

Lasse Hjort Jakobsen
Employment: Novo Nordisk
Honoraria: Takeda, Roche

No other potential conflicts of interest were reported.

REFERENCES

1. Eichenauer DA, Aleman BMP, André M, et al: Hodgkin lymphoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 29:iv19-iv29, 2018
2. Swerdlow SH: WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues Revised 4. Lyon, International Agency for Research on Cancer, 2017
3. Biccler JL, Glimelius I, Eloranta S, et al: Relapse risk and loss of lifetime after modern combined modality treatment of young patients with Hodgkin lymphoma: A Nordic Lymphoma Epidemiology Group Study. J Clin Oncol 37:703-713, 2019
4. Hasenclever D, Diehl V, Armitage JO, et al: A prognostic score for advanced Hodgkin's disease: International Prognostic Factors Project on Advanced Hodgkin's Disease. N Engl J Med 339: 1506-1514, 1998
5. Diefenbach CS, Li H, Hong F, et al: Evaluation of the International Prognostic Score (IPS-7) and a Simpler Prognostic Score (IPS-3) for advanced Hodgkin lymphoma in the modern era. Br J Haematol 171:530-538, 2015
6. Hayden AR, Lee DG, Villa D, et al: Validation of a simplified International Prognostic Score (IPS-3) in patients with advanced-stage classic Hodgkin lymphoma. Br J Haematol 189:122-127, 2020
7. Biccler JL, El-Galaly TC, Bøgsted M, et al: Clinical prognostic scores are poor predictors of overall survival in various types of malignant lymphomas. Leuk Lymphoma 60:1580-1583, 2019
8. Rodday AM, Parsons SK, Upshaw JN, et al: The advanced-stage Hodgkin lymphoma international prognostic index: Development and validation of a clinical prediction model from the HoLISTIC Consortium. J Clin Oncol 41:2076-2086, 2023
9. Biccler JL, Eloranta S, de Nully Brown P, et al: Optimizing outcome prediction in diffuse large B-cell lymphoma by use of machine learning and nationwide lymphoma registries: A Nordic Lymphoma Group Study. JCO Clin Cancer Inform 2:1-13, 2018
10. Ghasemieh A, Lloyed A, Bahrami P, et al: A novel machine learning model with stacking ensemble learner for predicting emergency readmission of heart-disease patients. Decis Anal J 7:100242, 2023
11. Arboe B, El-Galaly TC, Clausen MR, et al: The Danish National Lymphoma Registry: Coverage and data quality. PLoS One 11:e0157999, 2016
12. RKKP, Dansk Lymfom Database, Dansk Kronisk Lymfatisk Leukæmi Database. Malignt Lymfom og CLL National årsrapport. https://www.sundhed.dk/content/cms/89/4689_lyfo-cll-aarsrapport-2022.pdf
13. Ekberg S, Smedby KE, Glimelius I, et al: Trends in the prevalence, incidence and survival of non-Hodgkin lymphoma subtypes during the 21st century—A Swedish Lymphoma Register Study. Br J Haematol 189:1083-1092, 2020

14. Ellin F, Landström J, Jerkeman M, et al: Real-world data on prognostic factors and treatment in peripheral T-cell lymphomas: A study from the Swedish Lymphoma Registry. Blood 124:1570-1577, 2014
15. Larsen IK, Småstuen M, Johannesen TB, et al: Data quality at the Cancer Registry of Norway: An overview of comparability, completeness, validity and timeliness. Eur J Cancer 45:1218-1231, 2009
16. Wey A, Connett J, Rudser K: Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. Biostatistics 16:537-549, 2015
17. Van Der Laan MJ, Polley EC, Hubbard AE: Super learner. Stat Appl Genet Mol Biol 6:Article25, 2007
18. Graf E, Schmoor C, Sauerbrei W, et al: Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18:2529-2545, 1999
19. Simon N, Friedman J, Hastie T, et al: Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 39:1-13, 2011
20. Mardia K, Kent J, Bibby J: Multivariate Analysis (ed 1). London, UK, Academic Press, 1979. https://www.elsevier.com/books/multivariate-analysis/mardia/978-0-08-057047-1
21. Royston P, Lambert PC: Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. College Station, TX, Stata Press, 2011
22. Steyerberg EW: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating/Elektronis. Cham, Springer International Publishing, 2019
23. Blanche P, Latouche A, Viallon V: Time-Dependent AUC With Right-Censored Data: A Survey Study, in Lee ML., Gail M, Pfeiffer R, et al (eds), Risk Assessment and Evaluation of Predictions (Lecture Notes in Statistics, vol 215). New York, NY, Springer, 2013, pp. 239-251
24. Blanche P, Dartigues JF, Jacqmin-Gadda H: Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 32:5381-5397, 2013
25. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30:1105-1117, 2011
26. Harrell FE Jr, Califf RM, Pryor DB, et al: Evaluating the yield of medical tests. JAMA 247:2543-2546, 1982
27. Varma S, Simon R: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91, 2006
28. Steyerberg EW, Vergouwe Y: Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Eur Heart J 35:1925-1931, 2014
29. Schemper M, Smith TL: A note on quantifying follow-up in studies of failure time. Control Clin Trials 17:343-346, 1996
30. Lockney NA, Yang JC: Radiation therapy for advanced-stage Hodgkin lymphoma. Adv Radiat Oncol 5:809-816, 2020
31. André MPE, Girinsky T, Federico M, et al: Early positron emission tomography response-adapted treatment in stage I and II Hodgkin lymphoma: Final results of the randomized EORTC/LYSA/FIL H10 trial. J Clin Oncol 35:1786-1794, 2017
32. Engert A, Haverkamp H, Kobe C, et al: Reduced-intensity chemotherapy and PET-guided radiotherapy in patients with advanced stage Hodgkin's lymphoma (HD15 trial): A randomised, open-label, phase 3 non-inferiority trial. Lancet 379:1791-1799, 2012
33. Borchmann P, Plütschow A, Kobe C, et al: PET-guided omission of radiotherapy in early-stage unfavourable Hodgkin lymphoma (GHSG HD17): A multicentre, open-label, randomised, phase 3 trial. Lancet Oncol 22:223-234, 2021
34. Johnson P, Federico M, Kirkwood A, et al: Adapted treatment guided by Interim PET-CT scan in advanced Hodgkin's lymphoma. N Engl J Med 374:2419-2429, 2016
35. Borchmann P, Goergen H, Kobe C, et al: PET-guided treatment in patients with advanced-stage Hodgkin's lymphoma (HD18): Final results of an open-label, international, randomised phase 3 trial by the German Hodgkin Study Group. Lancet 390:2790-2802, 2017
36. Biccler JL, El-Galaly TC, Bøgsted M, et al: Clinical prognostic scores are poor predictors of overall survival in various types of malignant lymphomas. Leuk Lymphoma 60:1580-1583, 2019
37. Banerjee D: Recent advances in the pathobiology of Hodgkin's lymphoma: Potential impact on diagnostic, predictive, and therapeutic strategies. Adv Hematol 2011:439456, 2011
38. Jachimowicz RD, Klapper W, Glehr G, et al: Gene expression-based outcome prediction in advanced stage classical Hodgkin lymphoma treated with BEACOPP. Leukemia 35:3589-3593, 2021
39. Sánchez-Espiridión B, Montalbán C, López A, et al: A molecular risk score based on 4 functional pathways for advanced classical Hodgkin lymphoma. Blood 116:e12-e17, 2010
40. Eloranta S, Boman M: Predictive models for clinical decision making: Deep dives in practical machine learning. J Intern Med 292:278-295, 2022
41. Aldin A, Umlauff L, Estcourt LJ, et al: Interim PET-results for prognosis in adults with Hodgkin lymphoma: A systematic review and meta-analysis of prognostic factor studies. Cochrane Database Syst Rev 1:CD012643, 2020