https://doi.org/10.1016/j.hpb.2018.10.003 HPB

ORIGINAL ARTICLE

Performance validation of the ALPPS risk model

Michael Linecker[1], Christoph Kuemmerli[1], Patryk Kambakamba[1], Andrea Schlegel[2], Paolo Muiesan[2], Ivan Capobianco[3], Silvio Nadalin[3], Orlando J. Torres[4], Arianeb Mehrabi[5], Gregor A. Stavrou[6,7], Karl J. Oldhafer[7,8], Georg Lurje[9], Deniz Balci[10], Hauke Lang[11], Ricardo Robles-Campos[12], Roberto Hernandez-Alejandro[13,14], Massimo Malago[15], Eduardo De Santibanes[16], Pierre-Alain Clavien[1] & Henrik Petrowsky[1]

[1]Swiss HPB and Transplantation Center, Department of Surgery, University Hospital Zurich, Switzerland, [2]Liver Unit, Queen Elizabeth Hospital Birmingham, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK, [3]Department of General, Visceral and Transplant Surgery, University Hospital Tübingen, Tübingen, Germany, [4]Department of Surgery, Universidade Federal do Maranhão, Sao Luis, MA, Brazil, [5]Department of General, Visceral, and Transplantation Surgery, University of Heidelberg, Heidelberg, [6]Department of Abdominal, Thoracic and Pediatric Surgery, Saarbruecken General Hospital, Saarbruecken, [7]Semmelweis University, Budapest, Campus Hamburg, [8]Department of General and Abdominal Surgery, Asklepios Hospital Barmbek, Hamburg, [9]Department of Surgery and Transplantation, University Hospital RWTH Aachen, Aachen, Germany, [10]Department of Surgery, Ankara University, Ankara, Turkey, [11]Department of General, Visceral, and Transplant Surgery, Universitatsmedizin Mainz, Mainz, Germany, [12]Department of Surgery and Liver and Pancreas Transplantation, Virgen de la Arrixaca Clinic and University Hospital, Murcia, Spain, [13]Department of Surgery, Division of HPB Surgery and Liver Transplantation, London Health Sciences Centre, London, Ontario, Canada, [14]Division of Transplantation, Hepatobiliary Surgery, University of Rochester, Rochester, USA, [15]Department of HPB- and Liver Transplantation Surgery, University College London, Royal Free Hospitals, London, UK, and [16]Department of Surgery, Division of HPB Surgery, Liver Transplant Unit, Italian Hospital Buenos Aires, Argentina

Abstract

Background: Based on the International ALPPS registry, we have recently proposed two easily applicable risk models (pre-stage1 and 2) for predicting 90-day mortality in ALPPS but a validation of both models has not been performed yet.

Methods: The validation cohort (VC) was composed of subsequent cases of the ALPPS registry and cases of centers outside the ALPPS registry.

Results: The VC was composed of a total of 258 patients including 70 patients outside the ALPPS registry with 32 cases of early mortalities (12%). Development cohort (DC) and VC were comparable in terms of patient and surgery characteristics. The VC validated both models with an acceptable prediction for the pre-stage 1 (*c*-statistic 0.64, $P = 0.009$ *vs*. 0.77, $P < 0.001$) and a good prediction for the pre-stage 2 model (*c*-statistic 0.77, $P < 0.001$ *vs*. 0.85, $P < 0.001$) as compared to the DC. Overall model performance measured by Brier score was comparable between VC and DC for the pre-stage 1 (0.089 *vs*. 0.081) and pre-stage 2 model (0.079 *vs*. 0087).

Conclusion: The ALPPS risk score is a fully validated model to estimate the individual risk of patients undergoing ALPPS and to assist clinical decision making to avoid procedure-related early mortality after ALPPS.

Received 20 July 2018; accepted 3 October 2018

Correspondence

Henrik Petrowsky, Swiss HPB and Transplant Center Zurich, Department of Surgery and Transplantation, University Hospital Zürich, Rämistrasse 100, CH-8091 Zürich, Switzerland. E-mail: Henrik.Petrowsky@usz.ch

Introduction

ALPPS (Associating liver partition and portal ligation) is a new two-stage hepatectomy variant,[1,2] which has gained considerable interest in hepatobiliary centers over the last years.[3] The principle of this staged procedure is based on a combination of portal vein

Preliminary results of this work were presented at the 12th Biennial Congress of the European-African Hepato-Pancreato-Biliary Association (E-AHPBA) May 23–26, 2017 in Mainz, Germany.

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

occlusion and parenchymal transection at the first stage causing accelerated contralateral liver hypertrophy before completion hepatectomy at second stage. This feature is the major advantage of this procedure leading to a much higher resectability rate as compared to the conventional two-stage hepatectomy. The controversy of ALPPS versus two-stage hepatectomy has been recently addressed by the Scandinavian LIGRO trial where patients with colorectal liver metastases (CRLM) were randomized into either the ALPPS or portal vein occlusion (ligation or embolization) arm.[4] In this trial, ALPPS achieved an overall resectability rate of 92% compared to only 57% in the portal vein occlusion arm.[4] Today, CRLM can be considered as the leading indication for ALPPS particularly in situations of bilobar tumor involvement with insufficient future liver remnant volume. This also applies to situations of extensive instrumentation by ablation and/or resection of the future liver remnant and chemotherapy-injured liver. Tumor entities other than CRLM are still not very common indications for ALPPS and require further evaluation.[5]

A major criticism of ALPPS, especially in the pioneer phase of this procedure, was the initially experienced high rate of perioperative morbidity and mortality. However, the initial experience with data collection in the International ALPPS Registry has gradually led to the identification of various risk factors for early mortality after this procedure.[6–9] Based on these observations the ALPPS risk score was recently created from the international registry cohort in order to estimate and predict the 90-day or in-hospital mortality risk of individual patients either upfront before stage 1 or before stage 2 surgery.[10] The main purpose of this risk model was to provide an assisting tool for the hepatobiliary surgeon, which may help guiding treatment decisions whether to proceed with ALPPS or not. In such a situation, the predicted pre-stage 2 risk may assist the decision to proceed, delay, or even abort stage 2 surgery. How much risk is too high needs to be discussed case by case. The majority of clinicians would agree that a predicted mortality of beyond 50% would be too high to proceed with ALPPS.

As indicated at time of development of the ALPPS risk score, internal validation in the development cohort was not performed due to the limited number of events and the intention of future validation in a separate cohort.[10] Therefore, the current study was conducted as a follow-up project to validate the performance of the ALPPS risk score in a population, which was composed of mixed cases entered in the ALPPS registry after the development cohort and outside of the registry.

Material and methods

Study design

The primary goal of this study was to validate the previously published ALPPS risk prediction model before stage-1 and stage-2 surgery[10] in a multicenter cohort including patients in- and outside the ALPPS registry. Approval to enter patients into the international ALPPS Registry was obtained by the Cantonal Ethics Committee of Zurich (KEK 2013-0326). Data inclusion from centers outside the registry was permitted by local ethics approval. In addition, the registry study was registered at ClinicalTrials.gov (NCT01924741). The electronic data entry system used the specialized clinical trial software secuTrial® (Interactive System, Berlin, Germany). Data monitoring was ensured by the international ALPPS Registry coordination from the University Hospital Zurich in Switzerland. The Scientific Committee of the ALPPS Registry approved the present study on November 28, 2016 (http://www.alpps.net/?q=node/83). Data extraction for analysis was performed on November 09, 2017.

Definition of the development and validation cohort

For the development cohort (DC), the original database of the ALPPS risk score was used.[10] The validation cohort (VC) was composed of cases of high-volume ALPPS centers (5 cases per center) entered in the ALPPS registry after development of the ALPPS risk score (*temporal cohort*) and recruited outside the ALPPS registry (*external cohort*). The temporal cohort was composed of all patients of high-volume centers entered in the ALPPS registry after September 29, 2015, while those entered before were used for the development cohort. In accordance with the DC, only patients with data on 90-day follow-up were included in the analysis. Further inclusion criteria of the validation cohort were in accordance with those used in the DC.[10] Besides parameters directly available in the registry, composite parameters were calculated: The Model of End Stage Liver Disease (MELD) score was calculated by the formula MELD = 3.78 × ln (bilirubin in mg/dL) + 11.2 × ln (INR) + 9.57 × ln (creatinine in mg/dL) + 6.43.[11] Liver failure was defined by the International Study Group of Liver Surgery (ISGLS) criteria as previously reported.[8,12]

Calculation of the ALPPS risk score

For both DC and VC, the ALPPS pre-stage 1 and pre-stage 2 risk score and the associated mortality risk in percent was calculated for each patient.[10] The pre-stage 1 score is composed of the two parameters tumor type and age. The sum of the respective score points results in a certain risk of 90-day or in-hospital mortality (Table 1). The pre-stage 2 score contains the risk points of stage 1 and adds information on the interstage interval including major interstage complications (defined as Clavien Dindo grade $\geq$ 3b[13]) and the two laboratory parameters, serum bilirubin and creatinine. While the pre-stage 1 score is a categorical model, the pre-stage 2 score represents a continuous model using a risk formula (Table 1).

Statistical analysis

Descriptive statistics and univariate analyses

Categorical variables were expressed in absolute numbers and percent, whereas continuous variables were expressed in median and interquartile range (IQR) throughout the manuscript. Data completeness was checked for all variables and presented in

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

Table 1 ALPPS pre-stage 1 and pre-stage 2 risk score[10]

	Risk points
Pre-stage 1 variables	
Tumor type	
CRLM (*reference*)	0
Non-CRLM, non-biliary	1
Biliary	2
Age ≥ 67 years	3
Pre-stage 2 variables	
Pre-stage 1 score, per point	0.66
Inter-stage complications ≥3b	1.2
Serum bilirubin pre-stage 2	1.5
Serum creatinine pre-stage 2	1.7

Pre-stage-1 scores of 0, 1, 2, 3, 4, and 5 were associated with **early mortality risk** of 2.7%, 4.9%, 8.6%, 15%, 24%, and 37%.

Pre-stage 2 Score=

0.66 × (Pre-stage 1 score) +

1.2 × (1 = complications ≥3b; 0 = complications <3b) +

1.5 × $\log_{10}$ (10 × bilirubin pre-stage 2 in mg/dL) +

1.7 × $\log_{10}$ (10 × creatinine pre-stage 2 in mg/dL)

Mortality risk (%) = odds/(1 + odds)

odds = exp (–6.9 + pre-stage 2 risk score)

Abbreviations: ALPPS, Associating liver partition and portal vein ligation for staged hepatectomy; CRLM, colorectal liver metastases.

percent. For univariate comparisons the χ^2 and test was used for categorical, the Mann-Whitney-*U* and Kruskal–Wallis test for continuous variables when appropriate. *P* values ≤ 0.05 were considered statistically significant.

Validation analysis

Predictive performance of the pre-stage 1 and 2 model was tested in three steps comparing discrimination, calibration, and overall performance metrics of the DC and VC. *First*, discrimination, a measure of correct risk classification was tested by Receiver-Operating-Characteristic (ROC) curve analysis and discrimination slope. ROC curve analysis was performed for both models to test their discriminatory ability for the 90-day and/or in-hospital mortality. Concordance (*c*)-statistics and the statistical significance of prediction were used to compare the respective groups numerically. *P* values ≤ 0.05 were considered statistically significant. The discrimination slope was calculated as difference between the mean predicted probability of death in patients with and without 90-day and/or in-hospital mortality in the DC and the VC, respectively.[14] *Second*, calibration, a measure of comparing prediction with actual outcome was tested by the median difference of predicted and the respective observed (Δ P/O) rates of mortality (%). The mean predicted and actual mortality were plotted in calibration plots for low, intermediate, and high-risk groups. In the pre-stage 1 model the low risk groups had mortality rates of 2.7 and 4.9%, the intermediate risk group of 8.6 and 15%, and the high risk group of 24 and 37%, in the pre-stage 2 model mortality risks were grouped in <5%, 5–30%, and >30%, respectively. As a further measure of calibration, the Hosmer–Lemeshow test was applied as goodness-of-fit method[15] for both models in the DC and VC. *Third*, overall performance of the models was assessed using the Brier score as suggested.[14,16,17] In addition to assessing predictive performance, the optimal cut-off value for age, the variable with the highest discrimination was re-evaluated in the VC using accuracy analyses.[18]

All statistical analyses were performed using IBM SPSS Statistics version 24 (IBM Corporation, Armonk, NY) and Graph Pad Prism version 7 (GraphPad Software, Inc., La Jolla, CA).

Results

Study population

A total of 786 ALPPS patients were analyzed. The DC consisted of 528 patients from 38 centers, while the VC of 258 patients from 22 centers worldwide. Following the inclusion criteria of the ALPPS Risk Score[10] only centers who have included ≥5 cases and performed complete or partial parenchymal transection were included in the analysis. Other variants such as ablation-assisted transection or tourniquet ALPPS were excluded. The VC was composed of two sub-cohorts, a *temporal* (n = 188) and *external* (n = 70) VC. The temporal VC represents patients entered into the ALPPS Registry after the development of the ALPPS risk score, while patients recruited from 4 centers outside the ALPPS Registry built the external VC. Out of 422 patients who would have qualified for the VC, 164 (39%) had a missing primary endpoint and were therefore not included in the final analysis. To rule out a selection bias by patients with missing data, the groups with and without data on 90-day mortality were compared in terms of patient characteristics (Supplementary Table 1). There were no significant differences in demographic, tumor- and liver-related variables between both groups (Supplementary Table 1).

For both, temporal and external VC, the time of data acquisition after modeling of the ALPPS risk score was considered. This time did not necessarily reflect the year when these cases were performed. Data export for the ALPPS risk score development was performed on September 29, 2015[10] and 62%

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

(n = 159) of the VC were operated before this date. In other words, more than half of the cases used for validation were operated before the ALPPS risk score was developed. Failure to proceed to stage 2 occurred in 7% (n = 20) of cases of the VC and 2% (n = 11) of the DC.

Comparison of characteristics of the development and validation cohort

CRLM was the leading tumor entity in 66% of cases in the VC and 69% in the DC. Biliary tumors were represented in a slightly higher proportion in the VC (19%) as compared to the DC (15%). Non-colorectal/non-biliary tumors were comparable between both cohorts (15 *vs.* 16%). Mean age was 59 (50–67) years in the VC *versus* 62 (53–69) years in the DC ($P = 0.004$). Interestingly, liver baseline characteristics revealed a higher standardized future liver remnant (sFLR) in the VC (0.24 *vs.* 0.21, $P = 0.002$), but baseline bilirubin, INR, creatinine, and MELD score were not significantly different between both cohorts (Table 2).

Perioperative outcome of the development and validation cohort

Ninety-day and in-hospital mortality occurred in 12% (n = 32) of cases in the VC and in 9% (n = 47) in DC. Overall interstage complications were significantly higher in the VC (46 *vs.* 32%) but there was no difference in major complications (8 *vs.* 10%) (Table 3). The mean operation time of stage 1 was 250 minutes (180–356) in the VC and 305 minutes (250–393) in the DC. Length of hospital stay after stage 1 did not significantly differ between VC and DC (11 *vs.* 10 days). Liver volume before stage 2 surgery revealed a higher sFLR in the VC 0.41 (0.3–0.48) as compared to the DC 0.37 (0.30-0-45). Growth kinetics, however, including the absolute difference in sFLR (Δ sFLR) and sFLR increase showed no significant difference (0.17 *vs.* 0.15 and 76% vs. 66%). Serum bilirubin levels, INR, and creatinine were comparable in both groups. The MELD score showed a statistical difference although the effect size was very small between VC and DC, which indicates no clinical relevance. In contrast, liver failure rates according to the ISGLS criteria were significantly higher in VC (19%) compared to the DC (9%) (Table 3).

Predicted mortality risk in the development and validation cohort

Applying the risk model to all patients included, the median mortality risk pre-stage 1 consisting of the parameters age and tumor type was 2.7% (2.7–8.6) in the VC and 2.7% (2.7–15) in the DC. Pre-stage 2 risk, integrating the pre-stage 1 risk and adding complications after stage-1, serum bilirubin and creatinine levels before stage-2 was 2.4% (1.2–9.4) in the VC and 3.1% (1.4–11.3) in the DC (Table 2). Not unexpectedly, this data reflects that the majority of patients had low risks pre-stage-1 and pre-stage 2.

Table 2 Pre-stage 1 characteristics

Variable	Development cohort		Validation cohort		*P*
	n = 528	Data completion	n = 258	Data completion	
Demographics					
Age, years	62 (53–69)	95%	59 (50–67)	98%	0.007
Gender; male, n (%)	303 (59)	100%	160 (63)	99%	0.328
BMI, kg/m^2	25 (23–28)	94%	26 (23–28)	98%	0.929
Liver tumor		95%		95%	
CRLM, n (%)	343 (69)		154 (66)		0.452
Biliary tumors, n (%)	76 (15)		44 (19)		0.223
Non-CRLM/non-biliary, n (%)	78 (16)		36 (15)		0.932
Liver baseline performance					
sFLR pre-stage 1	0.21 (0.16–0.27)	85%	0.24 (0.18–0.29)	82%	0.002
Serum bilirubin, mg/dl	0.59 (0.40–0.90)	84%	0.60 (0.41–0.90)	92%	0.664
INR	1.0 (1.0–1.1)	79%	1.0 (1.0–1.1)	90%	0.002
Serum creatinine, mg/dl	0.81 (0.70–0.96)	74%	0.79 (0.66–0.92)	90%	0.106
MELD score	7 (6–8)	66%	7 (6–8)	86%	0.006
Outcome					
90-day or in-hospital mortality, n (%)	47 (9)	100%	32 (12)	100%	0.125
Overall futility risk pre-stage-1, %	2.7 (2.7–15.0)	92%	2.7 (2.7–8.6)	88%	0.754
Overall futility risk pre-stage-2, %	3.1 (1.4–11.3)	71%	2.4 (1.2–9.4)	81%	0.261

Abbreviations: BMI, body mass index; CRLM, colorectal liver metastases; sFLR, standardized future liver remnant; INR, international normalized ratio; MELD, Model of End-stage Liver Disease. Continuous variables presented as median and interquartile range (IQR). Categorical variables presented as count and percent (%).

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

Table 3 Pre-stage 2 characteristics

Variable	Development cohort		Validation cohort		*P*
	n = 528	Data completion	n = 258	Data completion	
Characteristics of stage 1					
Overall complications, n (%)	129 (32)	77%	115 (46)	97%	<0.001
Major complications, n (%)*	45 (10)	85%	20 (8)	97%	0.373
Operation time stage 1, min	305 (250–393)	82%	250 (180–356)	92%	<0.001
Hospital stay after stage 1, days	10 (7–14)	81%	11 (7–19)	87%	0.371
Interstage performance					
Liver volume characteristics					
sFLR pre-stage 2	0.37 (0.30–0.45)	77%	0.41 (0.34–0.48)	82%	0.041
Δ sFLR**	0.15 (0.10–0.20)	73%	0.17 (0.12–0.24)	76%	0.124
sFLR increase, %	66 (42–101)	66%	76 (48–107)	76%	0.427
Liver tests pre-stage 2					
Serum bilirubin, mg/dl	0.76 (0.47–1.29)	82%	0.63 (0.40–1.06)	94%	0.081
INR	1.1 (1.0–1.2)	79%	1.1 (1.0–1.2)	91%	0.458
Serum creatinine, mg/dl	0.71 (0.60–0.91)	73%	0.71 0.59–0.86)	90%	0.321
MELD score	8 (7–10)	64%	8 (7–10)	87%	0.044
ISGLS, n (%)	39 (9)	79%	47 (19)	94%	<0.001
Interstage interval, days	11 (8–15)	74%	12 (9–15)	97%	0.030
Stage 2 surgery					
Operation time, min	150 (112–200)	68%	153 (100–210)	88%	0.634

Abbreviations: sFLR, standardized future liver remnant; INR, international normalized ratio; MELD, Model of End-stage Liver Disease; ISGLS, International Study Group of Liver Surgery criteria. Continuous variables presented as median and interquartile range (IQR). Categorical variables presented as count and percent (%).
*, defined as complications ≥3b; **, reflects absolute change of sFLR volume before and after stage 1 surgery.

Identifying best age cut-off in the development and validation cohort

Age has turned out to have the highest predictive ability of 90-day mortality among all variables in the pre-stage 1 and pre-stage 2 model with a regression coefficient of 1.735.[10] Accuracy analysis was performed to identify best cut-off values at which age the risk of experiencing a 90-day mortality is significantly increased. The accuracy plot revealed an ideal cut-off of 67 years in the DC, and 64 years in the VC (Fig. 1). Pooling both cohorts (VC and DC) a best cut-off was achieved at 66 years.

Discriminatory ability of the ALPPS risk model in the development and validation cohort

ROC curve analysis of the pre-stage 1 model in the VC compared to the DC revealed an acceptable prediction for the pre-stage 1 model (*c*-statistic 0.64, $P = 0.009$ *vs.* 0.77, $P < 0.001$) and a good prediction for the pre-stage 2 model (*c*-statistic 0.77, $P < 0.001$ *vs.* 0.85, $P < 0.001$) (Fig. 2). Both models lost predictive ability when applied to the VC, with a *c*-statistic difference of 0.13 in the pre-stage 1 and 0.08 in the pre-stage 2 model. In addition to ROC curve analysis, the discrimination slope was used as a measure to separate prediction of patients with and without mortality

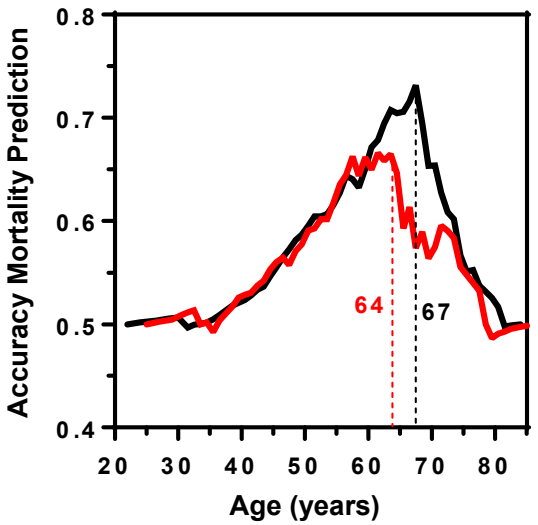

Figure 1 Accuracy analysis of determining the best age cut-offs in DC and VC. Accuracy of mortality prediction in relation to age (years) was calculated for all patients in the DC and VC and plotted to define its distribution. The black line represents the DC, the red line the VC. The age cut-off of 67 years to predict mortality as previously described[10] (DC) is comparable to the age cut-off of 64 years found in the VC

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

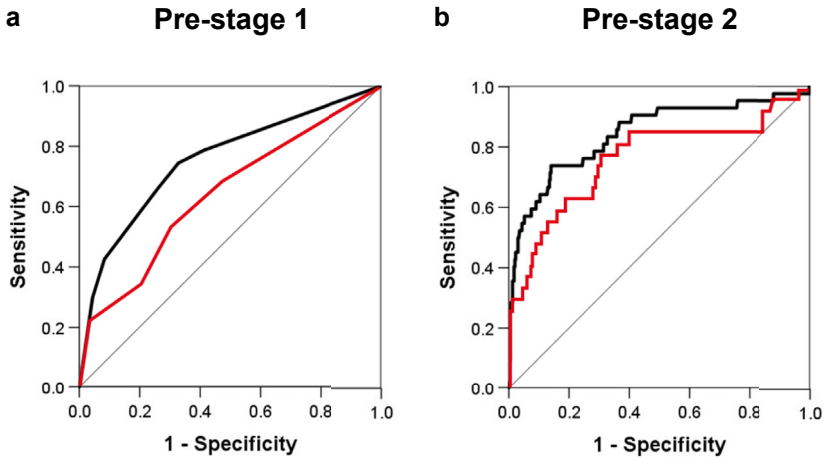

Figure 2 ROC curve analysis of pre-stage 1 and 2 prediction in DC and VC. ROC curve analysis of the ALPPS pre-stage 1 and pre-stage 2 model in the DC (n = 528; Panel a) as calculated in the ALPPS Risk Score[10] compared to the VC (n = 258; Panel b). Both prediction models revealed a significant discrimination of 90-day mortality in the DC and VC. A *c*-statistic of 0.77 ($P < 0.001$) *vs*. 0.64 ($P = 0.009$) was achieved in the pre-stage 1 model, and 0.77 ($P < 0.001$) *vs*. 0.85 ($P < 0.001$) in the pre-stage 2 model. The black lines represent the DC, the red lines the VC

(Fig. 3).[14] In the DC, the mean predicted mortality of patients with "no 90-day mortality" and "90-day mortality" was 7.3% and 25.6% resulting in a discrimination slope of 18.3%. Accordingly, the discrimination slope of the VC was 18.6% with a mean predicted mortality of 6.2% for the "no 90-day mortality" group and 24.8% for the "90-day mortality" group.

Calibration of the ALPPS risk model in the development and validation cohort

Comparing predicted *vs*. actual outcome is commonly referred as calibration of a model. In both risk models a continuous increase in actual mortality rates with increasing predicted risk was observed (Fig. 4). For simplification, patients were grouped in low, intermediate, and high-risk groups. In the pre-stage 1 model, low risk was defined by risk scores of 0 and 1 (2.7 and 4.9% predicted mortality), intermediate risk by 2 and 3 (8.6 and 15% predicted mortality), and high risk by 4 and 5 (24 and 37% predicted mortality). Predicted mortality risks of <5%, 5–30%, and >30% were assigned in the pre-stage 2 model for the respective risk categories low, intermediate, and high (Table 4).

Cases distribution of the low (63 *vs*. 64%), intermediate (26 *vs*. 25%), and high (11 *vs*. 11%) risk groups was comparable between the DC and VC in the pre-stage 1 model (Table 4, Fig. 2,

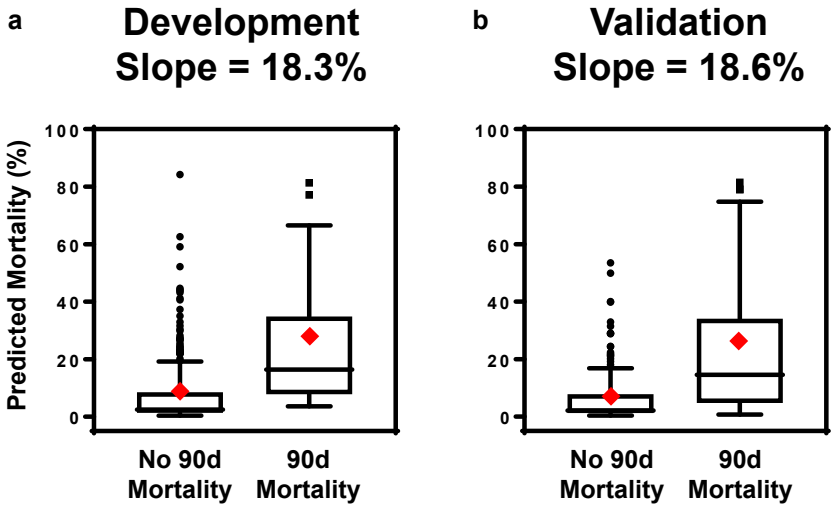

Figure 3 Discrimination slope of the DC and VD in the pre-stage 2 score. Box plots of predicted mortalities separated in groups with and without 90d-mortality for DC and VC in the pre-stage 2 model. The red squares represent the mean predicted 90d-mortality for the respective groups with 7.3 and 25.6% in the DC and 6.2 and 24.8% in the VC. The absolute difference in average predictions was 18.3% in the DC and 18.6% in the VC and called discrimination slope

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

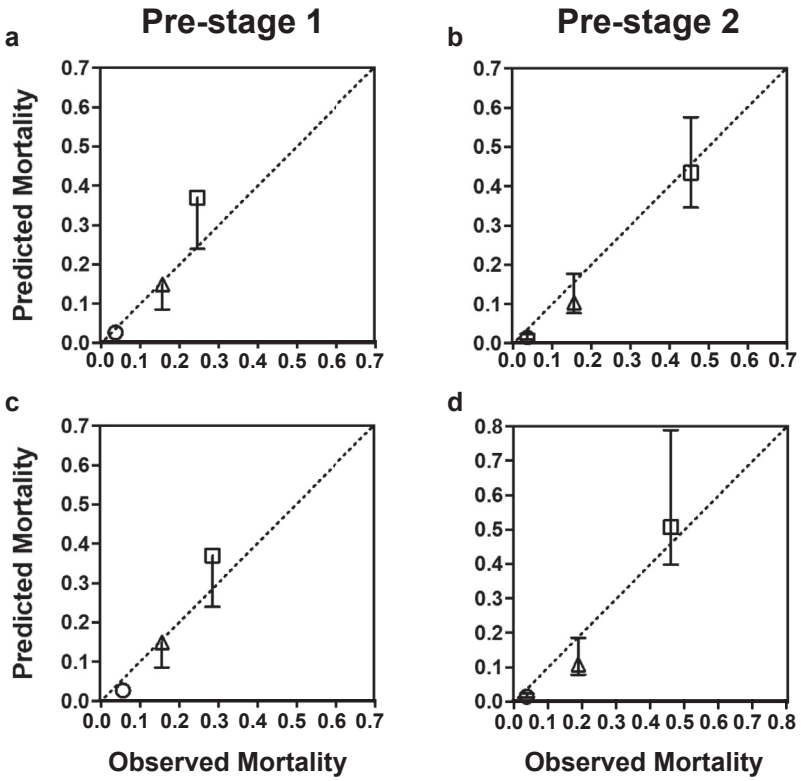

Figure 4 Calibration of pre-stage 1 and 2 models for DC and VC. Predicted mortality risks (median and interquartile range (IQR)) are plotted against observed mortality frequencies of three risk groups (*low*, *intermediate*, *high*). Open circle represents a *low-risk group*, which was defined with a mortality risk of 2.7 or 4.9% (0 or 1 risk points) in the pre-stage 1 model, and <5% in the continuous pre-stage 2 model. Open triangle represents an *intermediate-risk group* with a mortality risk of 8.6 or 15% (2 or 3 risk points) in the pre-stage 1 model, and a 5–30% mortality risk in the pre-stage 2 model. Open Square displays a *high-risk group* including a mortality risk of 24 or 37% (4 or 5 risk points) in the pre-stage 1 model, and >30% in the pre-stage 2 model. Figure panels a and b illustrate the DC, whereas panel c and d display the VC. The dotted oblique line shows an ideal line where mortality prediction is neither over- nor underestimated. As a result of a low number of high-risk cases both cohorts show a skewed distribution with only 11% of cases (DC, VC) in the pre-stage 1 model and 7 and 5% (DC, VC) in the pre-stage 2 model

Panel a,c). A similar distribution of low (60 *vs.* 64%), intermediate (33 *vs.* 31%), and high-risk (7 *vs.* 5%) cases was also observed for the pre-stage 2 model between the DC and VC (Table 4, Fig. 4, Panel b,d).

In the pre-stage 1 model, a median predicted mortality risk of 2.7% (low), 15% (intermediate), and 37% (high) was associated with an actual mortality of 4.2, 16, 25% in the DC and 4, 16, 25% in the VC. In the pre-stage 2 model, a median predicted mortality risk of 1.5% (low), 11% (intermediate), and 44% (high) was associated with an observed outcome of 4.0, 19, and 46% in the DC and a median predicted mortality risk of 1.5% (low), 11% (intermediate), and 50% (high) was associated with an actual outcome of 3.8, 19, and 46% in the VC. When DC and VC were compared, the differences between predicted and observed mortalities (Δ P/O) indicated no large deviations for the majority of the risk groups in both models (Table 4).

The Hosmer–Lemeshow goodness-of-fit method, which tests observed to predicted outcomes by decile of predicted probability was neither significant in the DC and VC in the pre-stage 1 ($P = 0.691$, $\chi^2 = 1.462$; $P = 0.291$, $\chi^2 = 3.742$) nor in the DC and VC in the pre-stage 2 model ($P = 0.711$, $\chi^2 = 5.431$; $P = 0.149$, $\chi^2 = 12.039$). This most likely indicates the exclusion of an overall calibration error.

Relation of risk category and individual model risk factors

The classification of patients into low, intermediate, and high risk followed the same pattern as of risk-determining variables in the DC and VC (Table 4). Median age of low, intermediate, and high-risk groups stepwise increased from 57, 69 to 73 years. In addition, a decrease of patients with CRLM, and an increase of patients with biliary tumors was noted from low to high-risk categories in both cohorts. A similar pattern was also observed for median serum bilirubin and creatinine pre-stage 2 in both cohorts, where a higher risk category was associated with higher serum bilirubin and creatinine levels (Table 4).

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

Table 4 Specific characteristics of risk categories

Variable	Development cohort (n = 528)				Validation cohort (n = 258)			
	Low*	Intermediate*	High*	*P*	Low**	Intermediate**	High**	*P*
Pre-stage 1 model								
No. of patients, n (%)	308 (63)	127 (26)	52 (11)		146 (64)	58 (25)	24 (11)	
Age, years	57 (48–62)	69 (65–73)	73 (69–76)	<0.001	55 (47–61)	68 (60–72)	72 (69–74)	<0.001
Tumor entity								
CRLM, n (%)	255 (76)	82 (24)	0 (0)	<0.001	121 (80)	31 (20)	0 (0)	<0.001
Biliary tumors, n (%)	0 (0)	45 (59)	31 (41)	<0.001	0 (0)	27 (63)	16 (37)	<0.001
Non-CRLM/non-biliary, n (%)	53 (72)	0 (0)	21 (28)	<0.001	25 (74)	1 (3)	8 (24)	0.001
Predicted mortality, %	2.7 (2.7–2.7)	15 (8.6–15)	37 (24–37)	<0.001	2.7 (2.7–2.7)	15 (8.6–15)	37 (14–37)	<0.001
Observed mortality %[#], n	4 (13)	16 (20)	25 (13)	<0.001	4 (13)	16 (20)	25 (13)	<0.001
Δ P/O, %	−1.3	−1	12		−1.3	−1	12	
Pre-stage 2 model								
No. of patients, n (%)	224 (60)	124 (33)	28 (7)		133 (64)	63 (30)	12 (6)	
Complications ≥3b, n (%)	7 (18)	22 (58)	9 (24)	<0.001	4 (29)	9 (64)	1 (7)	0.013
Bilirubin pre-stage-2, mg/dl	0.68 (0.40–1.10)	0.71 (0.50–1.45)	1.55 (0.89–4.34)	<0.001	0.53 (0.34–0.80)	0.80 (0.53–1.60)	3.05 (1.15–9.30)	<0.001
Creatinine pre-stage 2, mg/dl	0.70 (0.60–0.80)	0.79 (0.65–1.00)	0.97 (0.70–1.30)	<0.001	0.69 (0.55–0.80)	0.78 (0.60–0.95)	1.12 (0.76–1.51)	<0.001
Predicted mortality, %	1.5 (1.1–2.4)	11 (7.7–17)	44 (35–58)	<0.001	1.5 (1.1–2.3)	11 (7.7–18)	50 (35–78)	<0.001
Observed mortality,%[#], n	4 (9)	16 (20)	46 (13)	<0.001	4 (5)	19 (12)	46 (5)	<0.001
Δ P/O, %	−2.5	−5	−2		−2.5	−8	4	

Abbreviations: CRLM, colorectal liver metastases; P, predicted; O, observed.
*, defined by predicted mortality risks in the pre-stage 1 model: 0 or 1 risk points (low), 2 or 3 points (intermediate), and 4 or 5 points.
**, defined by predicted mortality risk in the pre-stage 2 model: <5% (low), 5–30% (intermediate), and >30% (high).
[#], refers to percent mortality per risk category. Continuous variables presented as median and interquartile range (IQR). Categorical variables presented as count and percent (%).

Overall performance of the pre-stage 1 and 2 model in the development and validation cohort

The Brier score is a well-established measure to test the overall performance of risk models for binary outcomes (mortality) by quantifying how close predictions are to the actual outcome.[14,16,17] This score integrates both discrimination and calibration. A Brier score of 0 represents a perfect prediction model while a Brier score of 0.25 indicates a non-informative model with a random chance of outcome. The Brier score of the pre-stage 1 model was 0.081 for the DC and 0.089 for the VC. The score of the pre-stage 2 model was 0.087 for the DC and 0.079 for the VC (Table 5).

Discussion

The present study represents the first validation of the previously published ALPPS risk model, which was created to predict the risk of 90-day and/or in-hospital mortality after ALPPS.[10] For both pre-stage 1 and pre-stage 2 models, the overall model performance in the VC was comparable to that in the DC, which was used for the development of the ALPPS risk score. Although the discriminatory ability was somewhat lower for the VC *versus* DC as well as the pre-stage 1 *versus* pre-stage 2 model, the calibration of both models revealed comparable outcomes between predicted and observed mortalities.

The validation of both ALPPS risk models is particularly challenging, as these models are subject of a very new procedure with a limited experience worldwide. The majority of cases are captured in the international ALPPS registry, which have been already used for the development of the risk score.[10] Therefore, a composite validation was chosen incorporating a temporal and external validation, which have been shown to serve as the most stringent validation strategy of a prognostic model.[19] Of note, "temporal" in this context reflects the date of data entry and not the date when the operation was performed as more than half of the cases in the VC were operated before the ALPPS risk score

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

Table 5 Characteristic performance measures of the pre-stage 1 und pre-stage 2 model for the development and validation cohort

Performance Measure	Pre-stage 1 Model		Pre-stage 2 Model	
	Development	Validation	Development	Validation
Overall				
Brier score[a]	0.081	0.089	0.087	0.079
Discrimination				
c-statistics[a]	0.77 ($p < 0.001$)	0.64 ($p = 0.009$)	0.85 ($p < 0.001$)	0.77 ($p < 0.001$)
Discrimination slope	n/a	n/a	18.3%	18.6%
Calibration				
Hosmer–Lemeshow test[b]	$X^2 = 1.46$, $p = 0.69$	$X^2 = 3.74$, $p = 0.29$	$X^2 = 5.43$, $p = 0.71$	$X^2 = 12.1$, $p = 0.15$

Abbreviation: n/a, not applicable.
[a] Statistics are scaled from 0 to 1. Lower Brier score and higher *c*-statistics represent better performance.
[b] Non-significant *p*-values represent better performance.

was created. Nevertheless, the present VC, which is composed of temporal and external cases, represents a data sample set, which is completely independent of the original data and model-fitting process. This type of validation is different from to the frequently used internal bootstrapping validation approach, which principally uses data of the DC.[19]

An important finding of this study is that the comparability of data samples in the DC and VC was given for the majority of pre-stage 1 and pre-stage 2 characteristics. However, small but significant differences were noted for the variables age, sFLR, operation time stage 1, and liver failure rate while others including MELD, INR, and length of interstage interval were significantly different but clinically irrelevant. Looking at the primary outcome variable of the ALPPS risk model, the 90-day mortality was 9% in the DC and 12% in the VC but without significant difference. The relatively high proportion of 12% in the VC is most likely related to the fact that more than half of the cases of the VC were operated before the ALPPS risk score was developed and, therefore, a significant number of cases fall in the pioneer phase of this procedure. However, these figures do not only represent the mortality rates of both cohorts but also demonstrate that the event rate was comparable in the DC and VC. Despite these minor differences, the case-mix of the development and validation cohort was overall comparable with each other.

At the development of the model, age has been identified as variable with the highest discriminatory ability amongst all other predictors.[10] It is well documented that continuous parameters always provide a better discriminative ability as compared to categorized continuous parameters using a cut-off. Accuracy analysis is able to determine the best cut-off value making this method very helpful to simplify prediction without losing too much discriminatory ability. We have applied this methodology at model development and identified age of 67 years as best discriminatory cut-off of 90-day mortality.[10] The VC revealed a slightly lower but similar cut-off of 64 years. This marginal difference of best age cut-offs might be related to the slightly lower median age of the VC as compared to the DC (59 vs. 62 years). However, the shapes of both accuracy curves were similar. The variable age seems to comprise many aspects that increase the risk of perioperative mortality. This includes the higher prevalence of co-morbidities, a globally slower regenerative response, and ultimately the failure-to-rescue in patients with advanced biological age.[20]

Prediction modeling is of major interest in clinical practice and particularly in surgery where the most accurate stratification of high and low-risk cases is of major relevance for clinical decision-making.[21–26] There is a growing body of prediction models which entered clinical practice to guide treatment decisions. Examples are the Framingham model predicting the 10-year risk of developing coronary heart disease,[27] the Gail's model for predicting the 10-year risk of developing breast cancer,[28] or the model of the National Surgical Quality Improvement Program (NSQIP) for predicting patient-individualized postoperative outcome.[29] The performance of these models have been evaluated in order to demonstrate valid risk prediction. Discrimination and calibration are two essential elements which are widely accepted to test the performance of prognostic prediction models.[14] In our analysis, we evaluated the performance of the pre-stage 1 and 2 model in the DC and VC by calculating model performance measures of discrimination, calibration, and overall performance. Although *c*-statistics metrics were somewhat lower in the VC than at development, the majority of performance measures were comparable between the VC and DC in both models. These findings indicate a consistent validation of the pre-stage 1 and pre-stage 2 ALPPS risk models.

All prognostic models have to face probabilities and uncertainties of predictions and are therefore vulnerable for certain degrees of misprediction. This is even documented in study populations, which are 20 times larger than our study population.[27] Therefore, it is of paramount importance to have valid and objective means of evaluating performance of risk prediction models. The discriminatory ability to predict outcome is a key element of modelling and was lower for the pre-stage 1

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003

compared to the pre-stage 2 model as well as for the VC compared to the DC. For instance, the *c*-statistics of the pre-stage 2 model of 0.77 in the VC indicates that prediction is correct in 77% of cases while incorrect in 23%. The lower *c*-statistics of the pre-stage 1 model in our validation study are most likely attributed to the fact that this model is determined by age and tumor entity while both variables are only one component among 3 others in the pre-stage 2 model. Therefore, the pre-stage 2 model is less dependent on the tumor entity as the pre-stage 1 model does. A future strategy to further enhance pre-stage 1 prediction might be the development of tumor-specific prediction models since risk factors of ALPPS in different tumor entities such as CRLM, HCC, or cholangiocarcinoma might be different. However, we have to keep in mind that even commonly used models such as the prediction model of coronary heart disease with more than 5000 patients had a *c*-statistics ranging from 0.7 to 0.8,[27] which would compare to discriminatory figures of the ALPPS risk model.

The strength of the present study is related to the multicenter setting combining external and temporal validation in a well-defined study population. Another important strength is the used systematic methodology of analyzing model performance at different levels including discrimination, calibration, and overall performance, which supported the validity of both models (Table 5). However, the present validation study is also associated with shortcomings, which are related to the relatively small validation cohort sample size of 258 patients consisting of less than half of patients of the DC. This is mainly attributed to the short period of time (1.5 years) since development of the ALPPS risk score. Inevitably, this lower sample size brings along a certain scattering of actual outcomes particularly in the high-risk group as a result of a low event rate in this risk category. Another limitation of the study is related to the nature of registry data. Although all cases with 90-day follow-up showed an acceptable data completeness in DC and VC, 39% of cases with missing 90-day follow-up could not be included in the present analysis. To prevent a selection bias, we have analyzed patient and liver baseline characteristics of patients included (with 90-day follow-up) and excluded (without 90-day follow-up) in the analysis (Supplementary Table 1). We demonstrated no statistical difference between both groups, which most likely excludes a selection bias. Missing data is a major problem not only of the international ALPPS registry, but of voluntary registries in general.[30] The global interest in ALPPS is exponentially growing, currently counting over 1000 cases in the international ALPPS registry (www.alpps.net). This rapid accumulation of cases frequently leads to data gaps. Critical evaluation of data completeness and validity of data entered as it is performed by the registry coordination on a regular basis is therefore essential. Despite these intense efforts, data entry of the participating centers cannot be enforced in a voluntary registry. A prospective analysis of further cases is welcome, aiming at confirmation and consolidation of the evidence provided by this cohort of patients.

In conclusion, the proposed ALPPS risk model[10] is a statistically validated tool to predict 90-day or in-hospital mortality in ALPPS either upfront or before stage 2 surgery. It has been designed to assist clinical decision making to avoid procedure-related early mortality after ALPPS. Decisions of denying ALPPS upfront or postponing stage 2 surgery in cases with high risk scores would avoid surgical intervention, which might be life span-reducing for these patients. However, recent experiences from ALPPS centers in and outside the registry indicate continuous improvement of safety of this procedure, which reached meanwhile mortality and morbidity outcome comparable to that accepted for major liver surgery.[31] A crucial next step is to identify patient populations, which may and may not benefit from ALPPS concerning tumor-related outcome.

Acknowledgement

The authors wish to thank Karin Petterson, Astrid Hirt and Christiane Nilles from the Department of Surgery, University Hospital Zurich for maintaining the International ALPPS Registry, as well as Lisette Paratore-Hari, PhD from the Clinical Trial Center (CTC) Zurich for managing the secuTrial® software of the ALPPS registry.

Financial support

The study was supported by the Clinical Research Priority Program of the University of Zurich as part of the project "Non-resectable liver tumors".

Conflict of interest

The authors declare no conflict of interest.

References

1. Schnitzbauer AA, Lang SA, Goessmann H, Nadalin S, Baumgart J, Farkas SA *et al*. (2012) Right portal vein ligation combined with in situ splitting induces rapid left lateral liver lobe hypertrophy enabling 2-staged extended right hepatic resection in small-for-size settings. *Ann Surg* 255:405–414.
2. de Santibanes E, Clavien PA. (2012) Playing Play-Doh to prevent post-operative liver failure: the "ALPPS" approach. *Ann Surg* 255:415–417.
3. Oldhafer KJ, Stavrou GA, van Gulik TM, Core Group. (2016) ALPPS-where do we stand, where do we go?: eight recommendations from the first international expert meeting. *Ann Surg* 263:839–841.
4. Sandström P, Røsok BI, Sparrelid E, Larsen PN, Larsson AL, Lindell G *et al*. (2018) ALPPS improves resectability compared with conventional two-stage hepatectomy in patients with advanced colorectal liver metastasis: results from a Scandinavian multicenter randomized controlled trial (LIGRO trial). *Ann Surg* 267:833–840.
5. Lang H, de Santibañes E, Schlitt HJ, Malagó M, van Gulik T, Machado MA *et al*. (2018 May, 1) 10th anniversary of ALPPS-lessons learned and quo vadis. *Ann Surg* [Epub ahead of print].
6. Truant S, Scatton O, Dokmak S, Regimbeau JM, Lucidi V, Laurent A *et al*. (2015) Associating liver partition and portal vein ligation for staged hepatectomy (ALPPS): impact of the inter-stages course on morbi-mortality and implications for management. *Eur J Surg Oncol* 41: 674–682.
7. Schadde E, Ardiles V, Robles-Campos R, Malago M, Machado M, Hernandez-Alejandro R *et al*. (2014) Early survival and safety of ALPPS: first report of the International ALPPS Registry. *Ann Surg* 260:829–836. discussion 836-8.

8. Schadde E, Raptis DA, Schnitzbauer AA, Ardiles V, Tschuor C, Lesurtel M *et al*. (2015) Prediction of mortality after ALPPS stage-1: an analysis of 320 patients from the international ALPPS registry. *Ann Surg* 262:780–786.
9. D'Haese JG, Neumann J, Weniger M, Pratschke S, Björnsson B, Ardiles V *et al*. (2016 Apr) Should ALPPS be used for liver resection in intermediate-stage HCC? *Ann Surg Oncol* 23:1335–1343.
10. Linecker M, Stavrou GA, Oldhafer KJ, Jenner RM, Seifert B, Lurje G *et al*. (2016) The ALPPS risk score: avoiding futile use of ALPPS. *Ann Surg* 264:763–771.
11. Wiesner R, Edwards E, Freeman R, Harper A, Kim R, Kamath P *et al*. (2003) Model for end-stage liver disease (MELD) and allocation of donor livers. *Gastroenterology* 124:91–96.
12. Rahbari NN, Garden OJ, Padbury R, Brooke-Smith M, Crawford M, Adam R *et al*. (2011) Posthepatectomy liver failure: a definition and grading by the international study group of liver surgery (ISGLS). *Surgery* 149:713–724.
13. Dindo D, Demartines N, Clavien PA. (2004) Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg* 240:205–213.
14. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N *et al*. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21:128–138.
15. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. (1997) A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 16:965–980.
16. B GW. (1950) Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78:1–3.
17. Gerds TA, Cai T, Schumacher M. (2008) The performance of risk prediction models. *Biom J* 50:457–479.
18. Fluss R, Faraggi D, Reiser B. (2005) Estimation of the Youden Index and its associated cutoff point. *Biom J* 47:458–472.
19. Altman DG, Royston P. (2000) What do we mean by validating a prognostic model? *Stat Med* 19:453–473.
20. Ghaferi AA, Dimick JB. (2016) Importance of teamwork, communication and culture on failure-to-rescue in the elderly. *Br J Surg* 103:e47–e51.
21. Shah N, Hamilton M. (2013) Clinical review: can we predict which patients are at risk of complications following surgery? *Crit Care* 17:226.
22. Lijftogt N, Luijnenburg TWF, Vahl AC, Wilschut ED, Leijdekkers VJ, Fiocco MF *et al*. (2017) Systematic review of mortality risk prediction models in the era of endovascular abdominal aortic aneurysm surgery. *Br J Surg* 104:964–976.
23. Sullivan PG, Wallach JD, Ioannidis JP. (2016) Meta-analysis comparing established risk prediction models (EuroSCORE II, STS score, and ACEF score) for perioperative mortality during cardiac surgery. *Am J Cardiol* 118:1574–1582.
24. Marufu TC, Mannings A, Moppett IK. (2015) Risk scoring models for predicting peri-operative morbidity and mortality in people with fragility hip fractures: qualitative systematic review. *Inj Int J Care Inj* 46:2325–2334.
25. Warnell I, Chincholkar M, Eccles M. (2015) Predicting perioperative mortality after oesophagectomy: a systematic review of performance and methods of multivariate models. *Br J Anaesth* 114:32–43.
26. Ganai S, Ferguson MK. (2013) Can we predict morbidity and mortality before an operation? *Thorac Surg Clin* 23:287–299.
27. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. (1998) Prediction of coronary heart disease using risk factor categories. *Circulation* 97:1837–1847.
28. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C *et al*. (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 81:1879–1886.
29. https://www.facs.org/quality-programs/acs-nsqip/about, accessed June, 2018.
30. Arts DG, De Keizer NF, Scheffer GJ. (2002) Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inf Assoc* 9:600–611.
31. Linecker M, Björnsson B, Stavrou GA, Oldhafer KJ, Lurje G, Neumann U *et al*. (2017) Risk adjustment in ALPPS is associated with a dramatic decrease in early mortality and morbidity. *Ann Surg* 266:779–786.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.hpb.2018.10.003.

Please cite this article in press as: Linecker M, et al., Performance validation of the ALPPS risk model, HPB (2018), https://doi.org/10.1016/j.hpb.2018.10.003