Journal of Clinical Epidemiology 121 (2020) 71−80

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Optimal sampling in derivation studies was associated with improved discrimination in external validation for heart failure prognostic models

Naotsugu Iwakami[a,b,c], Toshiyuki Nagai[a,d,*], Toshiaki A. Furukawa[c], Aran Tajika[c], Akira Onishi[c,e], Kunihiro Nishimura[f], Soshiro Ogata[g], Michikazu Nakai[g], Misa Takegami[f], Hiroki Nakano[a], Yohei Kawasaki[h], Ana Carolina Alba[i], Gordon Henry Guyatt[j], Yasuyuki Shiraishi[k], Shun Kohsaka[k], Takashi Kohno[k], Ayumi Goda[l], Atsushi Mizuno[m], Tsutomu Yoshikawa[n], Toshihisa Anzai[a,d], on behalf of the investigators for the WET-NaDEF Collaboration Project

[a]*Department of Cardiovascular Medicine, National Cerebral and Cardiovascular Center, Osaka, Japan*
[b]*Department of Research Promotion and Management, National Cerebral and Cardiovascular Center, Osaka, Japan*
[c]*Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/Public Health, Kyoto, Japan*
[d]*Department of Cardiovascular Medicine, Hokkaido University Graduate School of Medicine, Hokkaido, Japan*
[e]*Department of Rheumatology and Clinical Immunology, Kobe University Graduate School of Medicine, Kobe, Japan*
[f]*Department of Preventive Medicine and Epidemiology Informatics, National Cerebral and Cardiovascular Center, Osaka, Japan*
[g]*Center for Cerebral and Cardiovascular Disease Information, National Cerebral and Cardiovascular Center, Osaka, Japan*
[h]*Clinical Research Center, Chiba University Hospital, Chiba, Japan*
[i]*Division of Cardiology, Department of Medicine, University of Toronto, Toronto, Ontario, Canada*
[j]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada*
[k]*Division of Cardiology, Department of Medicine, Keio University School of Medicine, Tokyo, Japan*
[l]*Department of Cardiology, Kyorin University School of Medicine, Tokyo, Japan*
[m]*Department of Cardiology, St. Luke's International Hospital, Tokyo, Japan*
[n]*Division of Cardiology, Sakakibara Heart Institute, Tokyo, Japan*

Accepted 21 January 2020; Published online 29 January 2020

Abstract

Objectives: The objective of the study was to identify determinants of external validity of prognostic models.

Study Design and Setting: We systematically searched for studies reporting prognostic models of heart failure (HF) and examined their performance for predicting 30-day death in a cohort of consecutive 3,452 acute HF patients. We applied published critical appraisal tools and examined whether bias or other characteristics of original derivation studies determined model performance.

Results: We identified 224 models from 6,354 eligible studies. The mean c-statistic in the cohort was 0.64 (standard deviation, 0.07). In univariable analyses, only optimal sampling assessed by an adequate and valid description of the sampling frame and recruitment details to collect the population of interest (total score range: 0−2, higher scores indicating lower risk of bias) was associated with high performance (standardized $\beta = 0.25$, 95% CI: 0.12 to 0.38, $P < 0.001$). It was still significant after adjustment for relevant study characteristics, such as data source, scale of study, stage of illness, and study year (standardized $\beta = 0.24$, 95% CI: 0.07 to 0.40, $P = 0.01$).

Funding Statement: The WET-NaDEF collaboration project was supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology (Japan Society for the Promotion of Science [JSPS KAKENHI]), in Tokyo, Japan, Grant 23591062 and 26461088 (Dr. Yoshikawa); Grants-in-Aid for Young Scientists from JSPS KAKENHI, Grant 15K19402 (Dr. Nagai) and 18K15860 (Dr. Shiraishi); a Japan Health Labour Sciences Research, in Tokyo, Japan, Grant 14528506 (Dr. Yoshikawa); and the Sakakibara Clinical Research Grant for Promotion of Sciences, Japan, 2012, 2013, and 2014 (Dr. Yoshikawa); a grant from the Japan Agency for Medical Research and Development, in Tokyo, Japan, Grant 201439013C (Dr. Kohsaka), and a grant from the Japan Cardiovascular Research Foundation, in Tokyo, Japan, Grant 24-4-2 (Dr. Anzai). The funders played no role in conducting the research.

Disclosures: Dr. Furukawa reported personal fees from Meiji Seika, grants and personal fees from Mitsubishi-Tanabe, personal fees from MSD, personal fees from Pfizer, outside the submitted work. Dr. Kohsaka reported grants and personal fees from Bayer Yakuhin, grants from Daiichi Sankyo, personal fees from Bristol-Myers Squibb/Pfizer, outside the submitted work. All the other authors reported that they have no relationships relevant to the contents of this paper to disclose.

* Corresponding author. Department of Cardiovascular Medicine, Hokkaido University Graduate School of Medicine, Kita 15, Nishi 7, Kita-ku, Sapporo, Hokkaido 060-8638, Japan. Tel.: +81-11-706-6974; fax: +81-11-706-7874.

E-mail address: tnagai@huhp.hokudai.ac.jp (T. Nagai).

https://doi.org/10.1016/j.jclinepi.2020.01.011

Conclusion: Optimal sampling representing the gap between the population of interest and the studied population in derivation studies was a key determinant of external validity of HF prognostic models.

Keywords: Heart failure; Mortality; Prognostic model; Systematic review; Prognosis; Prediction; External validation; Study bias

1. Introduction

Heart failure (HF) is a leading cause of mortality and morbidity and a heavy social and economic burden in affluent societies [1]. In addition to innovative therapies, accurate prognostic assessment tools with optimal risk management within the existing framework of medicine are a major key to reduce these burdens [2]. Notably, they support appropriate decision-making by health care stakeholders, facilitating well-planned life management for patients themselves and efficient and effective patient and hospital management for health care providers and policymakers, using limited time and resources [2−5]. Inappropriate selection and application of prognostic models may conversely cause huge loss in terms of risk-treatment mismatch and regulatory failures [6−8]. The demand for efficient prediction models is higher than ever in the era of precision medicine [9].

Despite the plethora of prognostic models for HF currently available [1], no confident guides exist regarding how to select models and which models to apply among them [10]. It is recommended to select models replicated in a number of studies and derived from cohorts similar to the population in question for the target outcomes [10−15]. However, it would be hard to find a model with complete correspondence because the backgrounds of prognostic models are more or less different from the situations to be applied in terms of patients' characteristics, prevalence of disease, incidence of adverse events, available treatments, outcomes to be predicted, time span of prediction, and intended moment of prediction. Therefore, understanding which differences affect and what determines the predictive value of prognostic models is required for efficient utilization of the models [10,12−14].

In addition to similarities in study characteristics, the quality of the derivation study is a potential determinant of external validity of prognostic models [14−17]. A biased study would have a distorted spectrum of the population, predictors and outcomes, deriving models composed of inappropriate variables with inappropriate weighting [18]. This is why a reporting guideline for prediction studies [19] and guides for systematic review of prediction models [14,20,21] emphasize study bias. However, the effect of study bias on predictive ability has not been systematically examined because of lack of a standard, systematic, and quantitative assessment tool.

Recently, an expert panel has developed a method for critical appraisal of prognostic research, the Quality In Prognosis Studies (QUIPS) [22]. A standardized format for identifying important clinical characteristics in prediction models has also been proposed (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies [CHARMS]) [20]. To address the issue, we used a cohort of 3,452 Japanese patients with acute HF and examined the performance of existing prognostic models of HF in predicting 30-day death after admission in the cohort. To identify factors associated with greater vs. less predictive power, we examined characteristics of the derivation study, including the study biases and similarities of study characteristics to those in the Japanese cohort.

2. Materials and methods

2.1. Overall study design

Our study involved four steps. First, we systematically identified all studies reporting prognostic models of HF. Second, we assessed the risk of bias and other study characteristics of the original derivation studies based on the recently published assessment tools. Third, we examined the performance of the models in an existing cohort of HF. Fourth, we examined associations between model derivation studies and their predictive power. The institutional review boards at National Cerebral and Cardiovascular Center and Kyoto University, where the analysis was conducted, approved the use of the cohort data for this study (M29-059, R1135). The research was conducted in accordance with the Declaration of Helsinki. Supplementary Material provides full details of the study methodology.

2.2. Criteria to select studies for review

We included prognostic studies of HF reporting statistical models of any modeling methods that provided enough details to apply them in a new cohort, such as all the predictors and coefficients in the models. Predictors in the models needed to be included in the validation cohort for replication. The study purpose could be development of prediction models or exploration of contributing factors (predictor-finding studies) [20]. The target populations were HF patients with current or prior symptoms regardless of the stage of illness, subtype, and etiology. The target outcomes were death, hospitalization, heart transplantation, heart assist device implantation, or their combination, regardless of the timing of the outcome measures.

2.3. Search strategy for identification of studies

We searched studies using PubMed (1950 to May 2017). Supplementary Methods provide the detailed search

What is new?

Key findings

- We found that optimal sampling was the key determinant of the performance of prognostic models for heart failure (HF) among all the bias components and study characteristics in the derivation studies when we directly compared the 224 systematically identified models in a cohort of acute HF.

What this adds to what was known?

- When applying prognostic models, it has been conventionally recommended to select models replicated in a number of studies and derived from cohorts similar to the population in question for the target outcomes. However, these conventional practices were not verified at least in this study.

What is the implication and what should change now?

- All users and developers of HF prognostic models—and perhaps prognostic studies in general—need to be more conscious of study bias, before jumping to "big data" or "artificial intelligence."
- Many prognostic models have been proposed especially in cardiovascular medicine, but what determines their successful replication in new populations has not been fully investigated. Further evidences through quantitative researches are required for the appropriate use and development of prognostic models in any medical field.

strategy. We applied the search filter developed by Ingui et al. for prognostic studies [23,24]. We also included the reference lists of relevant systematic reviews to identify additional studies missed from the electronic searches.

2.4. Selection of studies and models

Two reviewers (N.I. and H.N.) selected the studies and models among all the records identified through the search methods. The PRISMA flow diagram (Supplementary Fig. S1) summarizes decisions made in the process [25]. After selecting the studies, we further selected one model in each study to avoid a clustering effect. We followed the priority list set in advance, which gave preference to mortality over hospitalization for modeling outcomes and to logistic or survival analysis over other modeling methods (Supplementary Methods).

2.5. Assessment of risk of bias and data collection in included studies

Independent pairs from a team of seven reviewers (N.I., A.T., A.O., K.N., M.T., M.N., and S.O.) assessed the risk of bias in the included derivation studies using the criteria outlined in QUIPS [22]. QUIPS comprises six domains with 31 subdomain items: domain 1, study participation (six subdomains); domain 2, study attrition (five subdomains); domain 3, prognostic factor measurement (six subdomains); domain 4, outcome measurement (three subdomains); domain 5, study confounding (seven subdomains); and domain 6, statistical analysis and reporting (four subdomains). Each subdomain item was answered as Yes (low risk of bias) or No (high risk of bias) and was scored as 1 and 0 for quantitative analysis.

We further extracted data of the study characteristics following CHARMS [20], referring to a reporting guideline for prediction models [19] and guides for systematic reviews [14,21], as well. CHARMS is a list of key items to extract from individual studies in a systematic review of prediction models.

2.6. Population used to apply identified models

We used cumulative data from the West Tokyo Heart Failure (WET-HF) registry and the National cerebral and cardiovascular center acute DEcompensated heart Failure (NaDEF) registry for applying identified prognostic models [4,26−28]. Both registries adopted the same eligibility criteria and recruited 3,781 consecutive acute HF patients admitted to six referral hospitals in East and West Japan. We excluded patients without vital status information since admission (n = 329) from analysis. The respective institutional review boards approved the study protocols of the two registries, and the study protocols have been registered at Japanese University Hospital Medical Information Network Clinical Trial Registration (UMIN000001171 and UMIN000017024, respectively).

2.7. Replication of identified models

We used clinical variables on admission to predict 30-day death after the index admission. We calculated the risk score for each patient for each included prognostic model and computed the c-statistic for each model. If the model was based on logistic regression, we calculated the risk scores as the sum of the variables weighted by the reported coefficients. If the model was survival analysis, we calculated the risk scores as the sum of the variables weighted by the logarithm of hazard ratios.

2.8. Statistical analyses

All continuous variables are shown as mean (standard deviation, SD) or median (interquartile range), as appropriate. The main analysis in this study was multivariable

linear regression analysis to identify the determinants of predictive value among study characteristics. The variables compared in the main analysis were the QUIPS [22] subdomain items for risk-of-bias assessment and the items in the CHARMS checklist [20] for other study characteristics including similarities of the cohorts. First, we examined the association between these items and c-statistics in the univariable analysis using unpaired *t*-test or analysis of variance for nominal and categorical variables, and simple linear regression analysis for continuous variables. We then included those significant items in the main analysis, considering the causal relationship, multicollinearity, and percentage of missing variables. Because some QUIPS subdomains are not necessarily independent, evaluating the same framework of bias from different aspects, we integrated similar subdomains to avoid multicollinearity by summating the score. We dichotomized multicategorical data in a prespecified clinically meaningful way. We checked the assumption of linearity and homoscedasticity between the predictors and outcomes by scatter plot. We performed all analyses using JMP Pro 13 (SAS Institute Inc., Cary, NC). All reported P values were two-sided, and the significance level was set at $P < 0.05$ adjusted for the number of multiple tests in the univariable analysis using the Bonferroni correction.

3. Results

3.1. Systematic review of prognostic models

Among 6,340 articles identified through electronic database search and 103 additional papers identified through reference list search, we identified 224 models in 224 studies (Supplementary Fig. S1). Supplementary Table S1 summarizes all included studies and their characteristics. Supplementary Fig. S2 summarizes the risk of bias according to QUIPS. The percentage agreement and the kappa statistic among the independent raters were 72% and 0.48, respectively, overall (Supplementary Table S2). Table 1 shows the characteristics of the derivation studies according to the CHARMS checklist [20]. Only 30 (13%) of the included studies reported c-statistics in the derivation cohort, the mean of which was 0.77 (SD, 0.06).

3.2. Replication and comparison of identified models

The identified models were applied to the cohort of 3,452 acute HF patients. Supplementary Table S3 shows the patient characteristics of the cohort. The outcome of 30-day death after admission occurred in 69 (2.0%).

The overall c-statistics of the identified models in the Japanese cohort had a mean of 0.64 (SD, 0.07) with approximately ½ between 0.6 and 0.7 and a quarter each between 0.5 and 0.6, and between 0.7 and 0.8 (Fig. 1). Supplementary Results list the c-statistics of the representative prognostic models for acute HF referred to in the clinical guideline [1].

Table 1 shows the association between the characteristics of the original derivation studies and their performance when applied to the Japanese cohort. Similarities of the cohorts in terms of the data source ($P = 0.02$), stage of illness ($P = 0.02$), age ($P = 0.002$), gender ($P = 0.003$), study year represented by the publication year ($P = 0.01$), and outcomes to be predicted ($P = 0.01$) were not statistically significant when the significance level was set at 0.05/50 = 0.001. Fig. 2 graphically shows the temporal improvement in performance of the prognostic models with an improvement from an average of 0.62 to 0.66 over approximately 20 years. Studies conducted in Japan did not present significantly higher c-statistics compared with those in other countries (c-statistic 0.67, SD 0.08 vs. 0.64, SD 0.08, $P = 0.06$).

Table 2 shows the correlation between the QUIPS assessment result and the predictive value of the models. Two subdomain items regarding the sampling frame and recruitment in the derivation studies were associated with higher c-statistic: "d. adequate description of the sampling frame and recruitment" ($P = 0.001$) and "e. adequate description of the period and place of recruitment" in QUIPS domain 1, study participation ($P = 0.003$). Both items exclusively measure the validity of participant recruitment and were conceptually and statistically inter-related (tetrachoric $\rho = 0.73$). The summative score of these two items (score range: 0 to 2, 2 indicating low risk of bias) was significantly associated with the c-statistic (standardized $\beta = 0.25$, 95% CI: 0.12 to 0.38, $P < 0.001$), even when adjusted for relevant study characteristics such as data source, scale of study, stage of illness, and study year (standardized $\beta = 0.24$, 95% CI: 0.07 to 0.40, $P = 0.01$) (Table 3). The results were robust in subgroup analysis excluding chronic HF studies (Supplementary Table S5).

4. Discussion

4.1. Major findings

In the present study, we demonstrated that optimal sampling was the key determinant of the performance of prognostic models for HF among all the bias components and study characteristics in the derivation studies. Similarities of derivation and replication cohorts showed association with predictive value in univariable analysis, such as data source, stage of illness, age, gender, study year, and outcomes to be predicted, but this methodological bias was the only significant. All users and developers of HF prognostic models—and perhaps prognostic studies in general—need to consider the apparent pre-eminent importance of this risk of bias item.

As we described in Supplementary Methods, we assessed not only if the sampling frame was described but also if it was valid. What we quantitatively measured in

Table 1. Characteristics of systematically identified studies and their impact on external validity of prognostic models

The study characteristic following the CHARMS checklist	Overall (n = 224)	Missing, n (%)	Averaged c-statistic or std β coefficient (95% CI)	P value
Source of data				
Cohort/registry	178 (79)	0	0.65 ± 0.07	0.02
Clinical trial	46 (21)		0.62 ± 0.07	
Prospective	179 (80)	0	0.64 ± 0.07	0.16
Retrospective	45 (20)		0.66 ± 0.07	
Participants				
Location				
Japan	23 (10)	0	0.67 ± 0.08	0.17
Europe	117 (52)		0.64 ± 0.07	
North America	39 (17)		0.64 ± 0.07	
Multi regions	20 (9)		0.64 ± 0.09	
Other regions	25 (11)		0.65 ± 0.07	
Number of centers				
Multicenter	115 (49)	0	0.64 ± 0.07	0.49
Single center	109 (47)		0.65 ± 0.07	
Setting				
Teaching hospitals	101 (43)	27 (12)	0.65 ± 0.07	0.69
Teaching and community hospitals	59 (25)		0.64 ± 0.07	
Community hospitals	19 (8)		0.64 ± 0.07	
Primary care	18 (8)		0.63 ± 0.09	
Stage of illness				
Acute heart failure	50 (22)	0	0.66 ± 0.08	0.02
Congestive heart failure	48 (21)		0.66 ± 0.07	
Chronic heart failure	126 (56)		0.63 ± 0.07	
Number of participants	513 [239, 1,611]	0		
Difference in number of participants[a]	2,962 [2,347, 3,236]		0.02 (−0.12 to 0.15)	0.81
Averaged age of participants, yrs	68 ± 7	4 (2)		
Difference in averaged age[a], yrs	6 [3, 11]		−0.21 (−0.34 to −0.08)	0.002
Male percentage of participants,%	64 ± 14	2 (1)		
Difference in male percentage[a],%	11 [7, 18]		−0.20 (−0.33 to −0.07)	0.003
Ischemic etiology percentage of participants, %	48 ± 18	69 (31)		
Difference in ischemic etiology percentage[a], %	24 [14, 32]		−0.10 (−0.25 to 0.06)	0.21
NYHA IV percentage of participants, %	8 [2, 32]	118 (53)		
Difference in NYHA IV percentage[a], %	30 [16, 35]		0.05 (−0.13 to 0.24)	0.58
Averaged SBP, mmHg	129 ± 12	86 (38)		
Difference in averaged SBP[a], mmHg	12 [6, 19]		−0.14 (−0.29 to 0.03)	0.10
Averaged LVEF of participants, %	36 ± 11	66 (29)		
Difference in averaged LVEF[a], %	11 [6, 17]		−0.13 (−0.28 to 0.03)	0.12
Study date: publication year	2,010 [2,007, 2,013]	0	0.17 (0.04 to 0.30)	0.01
Main outcomes to be predicted				
Death	173 (77)		0.65 ± 0.07	0.01
Death and hospitalization	41 (10)		0.64 ± 0.07	
Hospitalization	10 (4)		0.58 ± 0.05	
Outcome event percentage, %	27 [17, 38]	10 (4)		
Difference in outcome event percentage[a], %	25 [15, 36]		0.02 (−0.11 to 0.15)	0.81
Time span of prediction to outcome measurement, days	577 [365, 1,192]	12 (5)		
Difference in time span[a], days	547 [335, 1,162]		0.03 (−0.11 to 0.16)	0.67

(*Continued*)

Table 1. Continued

The study characteristic following the CHARMS checklist	Overall (n = 224)	Missing, n (%)	Averaged c-statistic or std β coefficient (95% CI)	P value
Candidate predictors				
Timing of predictor measurement				
At admission	37 (17)	34 (15)	0.66 ± 0.08	0.16
During hospital stay	13 (6)		0.66 ± 0.06	
At discharge	41 (18)		0.65 ± 0.08	
At patient presentation	99 (44)		0.63 ± 0.07	
Model development				
Modeling method				
Survival model	208 (93)	0	0.64 ± 0.07	0.07
Logistic regression model	14 (6)		0.66 ± 0.06	
Decision tree	2 (1)		0.56 ± 0.03	
Purpose of modeling				
Development of prognostic models	24 (11)	0	0.63 ± 0.07	0.60
To assess the impact of specific predictors	160 (71)		0.65 ± 0.07	
To assess the degree of contribution of each predictor	40 (18)		0.64 ± 0.07	

Abbreviations: CHARMS, CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies; CI, confidence interval; LVEF, left ventricular ejection fraction; NYHA, New York Heart Association; SBP, systolic blood pressure; SE, standard error; Std, standardized.

Values are mean ± SD, n (%) or median [interquartile range].

[a] Absolute difference in the variable between the derivation cohort and the external validation cohort.

the sampling bias score was the gap between the population of interest and the studied population in the derivation study. Those studies without valid presentation of the recruitment method are likely to have this gap. This gap was not necessarily apparent in the variables of the population characteristics. We suspect that the unrepresentativeness of the derivation cohort in the population under study would have brought inappropriate selection of variables with inappropriate weighting in the statistical models, resulting in inaccurate prognostic models [18].

Two representative systematic reviews regarding prognostic models for HF have been reported. Ouwerkerk et al. [29] demonstrated from meta-regression analysis of 117 HF prognostic models in 55 papers that the mortality models and registry-type studies showed higher c-statistics than hospitalization models and clinical trials, although stage of illness did not show significant difference. On the other hand, Rahimi et al. reviewed 64 models in 48 studies and reported that the mortality models showed higher c-statistics. They did not mention stage of illness, and the data source and study design were not related to the discriminative ability of the models [30]. Both studies used c-statistics in the derivation cohorts for their meta-analysis, whereas the present study reproduced each model in an

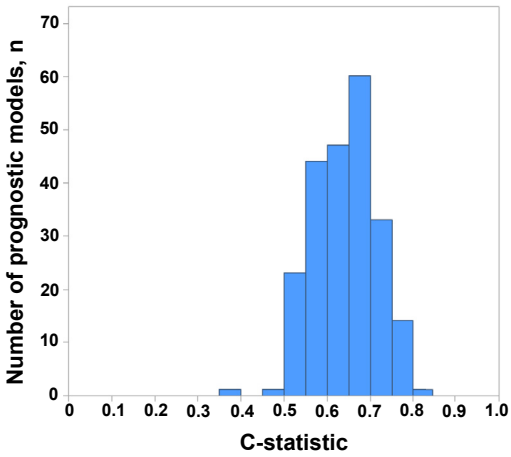

Fig. 1. Histogram of c-statistics of systematically identified prognostic models in the replication cohort.

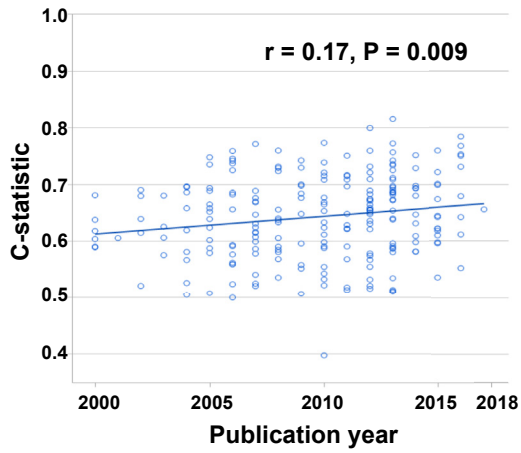

Fig. 2. Association between performance of prognostic models in the replication cohort and the publication year.

Table 2. The impact of study bias on external validity of identified prognostic models

Risk of bias item in QUIPS	Averaged c-statistic for low RoB [*n* (% out of 224)]	Averaged c-statistic for high RoB [*n* (% out of 224)]	*P* value
Domain 1. Study participation			
a. Adequate participation in the study by eligible persons	0.64 ± 0.07 [75 (33)]	0.64 ± 0.08 [149 (67)]	0.76
b. Description of the source population or population of interest	0.64 ± 0.07 [196 (88)]	0.63 ± 0.08 [28 (13)]	0.55
c. Description of the baseline study sample	0.65 ± 0.07 [94 (42)]	0.64 ± 0.07 [130 (58)]	0.44
d. Adequate and valid description of the sampling frame and recruitment	0.65 ± 0.07 [179 (80)]	0.61 ± 0.07 [45 (20)]	0.001
e. Adequate and valid description of the period and place of recruitment	0.65 ± 0.07 [134 (60)]	0.62 ± 0.07 [90 (40)]	0.003
f. Adequate and valid description of inclusion and exclusion criteria	0.65 ± 0.08 [102 (46)]	0.64 ± 0.07 [122 (54)]	0.15
Domain 2. Study attrition			
a. Adequate response rate for study participants	0.64 ± 0.07 [79 (35)]	0.65 ± 0.07 [145 (65)]	0.40
b. Description of attempts to collect information on participants who dropped out	0.64 ± 0.07 [79 (35)]	0.65 ± 0.07 [145 (65)]	0.40
c. Reasons for loss to follow-up are provided	0.64 ± 0.07 [79 (35)]	0.65 ± 0.07 [145 (65)]	0.40
d. Adequate description of participants lost to follow-up	0.63 ± 0.06 [10 (4)]	0.64 ± 0.07 [214 (96)]	0.47
e. There are no important differences between participants who completed the study and those who did not	0.63 ± 0.06 [10 (4)]	0.64 ± 0.07 [214 (96)]	0.47
Domain 3. Prognostic factor measurement			
a. A clear definition or description of the PF is provided	0.64 ± 0.07 [218 (97)]	0.66 ± 0.03 [6 (3)]	0.59
b. Method of PF measurement is adequately valid and reliable	0.65 ± 0.10 [210 (94)]	0.64 ± 0.07 [14 (6)]	0.53
c. Continuous variables are reported or appropriate cut points are used	0.64 ± 0.07 [147 (66)]	0.64 ± 0.08 [77 (34)]	0.40
d. The method and setting of measurement of PF is the same for all study participants	0.65 ± 0.06 [160 (71)]	0.63 ± 0.08 [64 (29)]	0.18
e. Adequate proportion of the study sample has complete data for the PF	0.63 ± 0.07 [74 (33)]	0.65 ± 0.07 [150 (67)]	0.14
f. Appropriate methods of imputation are used for missing PF data	0.63 ± 0.07 [74 (33)]	0.65 ± 0.07 [150 (67)]	0.14
Domain 4. Outcome measurement			
a. A clear definition of the outcome is provided	0.64 ± 0.07 [218 (97)]	0.63 ± 0.09 [6 (3)]	0.65
b. Method of outcome measurement used is adequately valid and reliable	0.65 ± 0.07 [125 (56)]	0.63 ± 0.07 [99 (44)]	0.01
c. The method and setting of outcome measurement is the same for all study participants	0.65 ± 0.07 [119 (53)]	0.64 ± 0.07 [105 (47)]	0.39
Domain 5. Study confounding			
a. All important confounders are measured	0.65 ± 0.07 [40 (18)]	0.64 ± 0.08 [118 (53)]	0.67
b. Clear definitions of the important confounders measured are provided	0.65 ± 0.07 [152 (68)]	0.61 ± 0.07 [6 (3)]	0.25
c. Measurement of all important confounders is adequately valid and reliable	0.65 ± 0.07 [154 (69)]	0.68 ± 0.07 [4 (2)]	0.32
d. The method and setting of confounding measurement are the same for all study participants	0.64 ± 0.08 [108 (48)]	0.65 ± 0.07 [50 (22)]	0.50
e. Appropriate methods are used if imputation is used for missing confounder data	0.62 ± 0.13 [7 (3)]	0.65 ± 0.07 [151 (67)]	0.38
f. Important potential confounders are accounted for in the study design or in the analysis	0.65 ± 0.07 [157 (70)]	– [1 (0)]	-
Domain 6. Statistical analysis and reporting			
a. Sufficient presentation of data to assess the adequacy of the analytic strategy	0.64 ± 0.07 [111 (50)]	0.65 ± 0.07 [113 (50)]	0.20

(*Continued*)

Table 2. Continued

Risk of bias item in QUIPS	Averaged c-statistic for low RoB [*n* (% out of 224)]	Averaged c-statistic for high RoB [*n* (% out of 224)]	*P* value
b. Strategy for model building is appropriate and is based on a conceptual framework or model	0.66 ± 0.07 [59 (26)]	0.64 ± 0.07 [165 (74)]	0.08
c. The selected statistical model is adequate for the design of the study	0.65 ± 0.07 [163 (73)]	0.63 ± 0.07 [61 (27)]	0.11
d. There is no selective reporting of results	0.64 ± 0.07 [188 (84)]	0.63 ± 0.08 [36 (16)]	0.38

Abbreviations: PF, prognostic factor; QUIPS, Quality In Prognosis Studies; RoB, risk of bias.
Values are mean ± SD.

external replication cohort. The quality of the derivation study was not examined in either study.

Ouwerkerk et al. [29] reported that the mean c-statistic of the identified models in derivation cohorts was 0.71. In the present study, the reported mean c-statistic in the derivation studies was 0.77. However, the c-statistics in Japanese cohort were as low as 0.64. This gap was not attributable to the misapplication of the models. Even if the models were limited to those for acute HF or those developed just for the purpose of prediction, the c-statistics were 0.66 and 0.63, respectively. Indeed, prognostic models for HF are not widely used in current real-world clinical practice [15]. This is partly due to the low predictive value of the prognostic models and party due to lack of treatment options depending on the prognosis. For better risk management of HF patients, development of optimal prognostic assessment methods along with development of innovative therapies are indispensable, and thus, better understanding of study bias is necessary for users and developers.

Our goal in the current investigation was to identify important steps to be taken in the development of prognostic models in derivation studies. We included predictor-finding studies and those to develop prognostic models because these studies have in common a similar process for deriving multivariable statistical models from cohorts [20,21]. The modeling purpose may affect selection and combination of variables in models; however, weighting of them is still under the effect of the study characteristics. Indeed, it did not cause a major difference in the performance of the models in our findings. We also included models regardless of the stage of illness because it was not necessarily distinguished in studies [31], and limiting studies to those with a clear statement on acute HF might cause bias. Surprisingly, stage of illness was not significantly related to the model performance, but the results were consistent with a previously reported review [29]. Even if chronic HF studies were excluded, the main results were robust.

We used QUIPS [22] and CHARMS [20] for critical appraisal of prognostic models. CHARMS offers data extraction items including risk-of-bias assessment and is intended exclusively for model development studies. On the other hand, QUIPS is an assessment tool for risk of bias in prognostic studies in general. It covers the risk-of-bias items in CHARMS and is more specialized and detailed for risk-of-bias assessment, although assessment of confounding is not applicable for model development studies. We included predictor-finding studies in the present study; hence, we used QUIPS for risk-of-bias assessment and CHARMS for data extraction of other study characteristics. We used QUIPS because the Prediction model Risk Of Bias ASsessment Tool (PROBAST) had not been published when we started our study in 2017 [32,33]. Although PROBAST contains several items tapping aspects necessary for

Table 3. Multivariable analysis to identify a determinant of external validity of prognostic models

Study characteristic variable	Std β coefficient (95% CI)	*P* value
Risk of bias in derivation studies		
QUIPS domain 1. Study participation d. Adequate and valid description of the sampling frame and recruitment and e. Adequate and valid description of the period and place of recruitment[a] (integrated score ranged from 0 to 2, 2 being low risk of bias)	0.24 (0.07 to 0.40)	0.01
Characteristics of derivation studies		
Source of data (nonclinical trial vs. clinical trial)	−0.02 (−0.19 to 0.16)	0.83
Source of data (prospective vs. retrospective)	−0.05 (−0.19 to 0.08)	0.44
Number of centers (single center vs. multicenter)	−0.06 (−0.21 to 0.09)	0.44
Stage of illness (acute heart failure vs. others)	−0.03 (−0.11 to 0.17)	0.68
Study date: publication year	0.11 (−0.02 to 0.25)	0.11

Abbreviations: QUIPS, Quality In Prognosis Studies; Std, standardized.
[a] Integrated score of two inter-related subdomain items, both exclusively evaluating study recruitment.

the development of prediction models, QUIPS allowed us to gain insights into several areas in more detail, such as sampling from the population.

4.2. Study limitations

The quality of a study is only assessable when reported in the article. It is only in recent years that the reporting of prediction models was formalized [19]. Study bias might not have been properly assessed if properly described in the articles. However, we suppose that those who were conscious of bias would report it and perform a less biased study and vice versa, and thus the result of the study would be robust. Second, the identified 224 studies might be a biased group among all prognostic studies in that they clarified all the necessary data for reproduction in the new cohort. The quality of the identified studies might be relatively high, which would reduce the association between study quality and the predictive value of prognostic models. Third, the kappa statistic was lower than expected. This is because QUIPS is a general bias assessment tool. Further studies are required to develop an assessment tool specialized for HF studies. Fourth, this study focused on the differences in discrimination but not in calibration among different prediction models, as the necessary information to calculate calibration measures was not available in most of the included studies especially in the predictor-finding studies. C-statistic is one of the representative measures of calibration; however, it is known to be insensitive to change. Finally, we only showed the result of one example of a combined acute HF cohort with one outcome in one disease entity, and thus, no universal conclusions about prognostic models in general can be drawn only by this study. Further studies are required for other outcomes, using other cohorts, in other medical fields along with appropriate feedback for future prediction studies.

5. Conclusion

Optimal sampling in derivation studies was a key determinant of the performance of HF prognostic models when applied in an acute HF cohort for predicting 30-day death after admission rather than similarities of characteristics in the studies. Consideration and presentation of study bias is important for all model users and developers.

CRediT authorship contribution statement

Naotsugu Iwakami: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing - original draft. **Toshiyuki Nagai:** Data curation, Funding acquisition, Investigation, Project administration, Writing - original draft. **Toshiaki A. Furukawa:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing - review & editing. **Aran Tajika:** Investigation, Writing - review & editing. **Akira Onishi:** Investigation, Writing - review & editing. **Kunihiro Nishimura:** Formal analysis, Investigation, Writing - review & editing. **Soshiro Ogata:** Investigation, Writing - review & editing. **Michikazu Nakai:** Investigation, Writing - review & editing. **Misa Takegami:** Investigation, Writing - review & editing. **Hiroki Nakano:** Investigation, Writing - review & editing. **Yohei Kawasaki:** Conceptualization, Methodology, Writing - review & editing. **Ana Carolina Alba:** Conceptualization, Methodology, Writing - review & editing. **Gordon Henry Guyatt:** Conceptualization, Methodology, Writing - review & editing. **Yasuyuki Shiraishi:** Investigation, Writing - review & editing. **Shun Kohsaka:** Funding acquisition, Investigation, Writing - review & editing. **Takashi Kohno:** Investigation, Writing - review & editing. **Ayumi Goda:** Investigation, Writing - review & editing. **Atsushi Mizuno:** Investigation, Writing - review & editing. **Tsutomu Yoshikawa:** Funding acquisition, Investigation, Writing - review & editing. **Toshihisa Anzai:** Funding acquisition, Investigation, Supervision, Writing - review & editing.

Acknowledgments

The authors are grateful for the contributions of all investigators, clinical research coordinators, and data managers involved in the WET and NaDEF study. The authors also thank Aya Ichizawa and Keiko Fujii (Kyoto University) for help collecting the full papers for the review.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.01.011.

References

[1] Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Drazner MH, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American heart association Task Force on practice guidelines. J Am Coll Cardiol 2013;62(16):e147−239.

[2] Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ 2013;346:e5595.

[3] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381.

[4] Shiraishi Y, Kohsaka S, Abe T, Mizuno A, Goda A, Izumi Y, et al. Validation of the get with the guideline-heart failure risk score in Japanese patients and the potential improvement of its discrimination ability by the inclusion of B-type natriuretic peptide level. Am Heart J 2016;171:33−9.

[5] Vogenberg FR. Predictive and prognostic models: implications for healthcare decision-making in a modern recession. Am Health Drug Benefits 2009;2(6):218−22.

[6] Lee DS, Tu JV, Juurlink DN, Alter DA, Ko DT, Austin PC, et al. Risk-treatment mismatch in the pharmacotherapy of heart failure. JAMA 2005;294:1240−7.
[7] Lee DS, Alba AC. Risks and benefits of risk prediction in acute heart failure. JACC Heart Fail 2015;3(10):748−50.
[8] Mortensen MB, Falk E. Limitations of the SCORE-guided European guidelines on cardiovascular disease prevention. Eur Heart J 2017;38:2259−63.
[9] Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793−5.
[10] Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140.
[11] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515−24.
[12] Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.
[13] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279−89.
[14] Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460.
[15] Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. JAMA 2018;320:27−8.
[16] Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;313:793−9.
[17] Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488−94.
[18] Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010;172:971−80.
[19] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55−63.
[20] Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11(10):e1001744.
[21] Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(5):1−12.
[22] Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013;158:280−6.
[23] Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc 2001;8:391−7.
[24] Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One 2012;7:e32844.
[25] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264−9. W64.
[26] Nagai T, Nishimura K, Honma T, Higashiyama A, Sugano Y, Nakai M, et al. Prognostic significance of endogenous erythropoietin in long-term outcome of patients with acute decompensated heart failure. Eur J Heart Fail 2016;18(7):803−13.
[27] Hamatani Y, Nagai T, Shiraishi Y, Kohsaka S, Nakai M, Nishimura K, et al. Long-term prognostic significance of plasma B-type natriuretic peptide level in patients with acute heart failure with reduced, mid-range, and preserved ejection fractions. Am J Cardiol 2018;121:731−8.
[28] Nagai T, Sundaram V, Shoaib A, Shiraishi Y, Kohsaka S, Rothnie KJ, et al. Validation of U.S. mortality prediction models for hospitalized heart failure in the United Kingdom and Japan. Eur J Heart Fail 2018;20(8):1179−90.
[29] Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. JACC Heart Fail 2014;2(5):429−36.
[30] Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, et al. Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail 2014;2(5):440−6.
[31] Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2016;18(8):891−975.
[32] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51−8.
[33] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1−33.