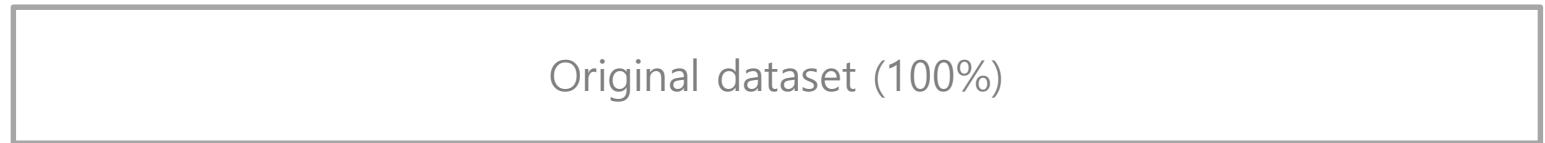




자료분할탭의 활용

■ 모형 평가를 위한 데이터셋 분할



Training set
(50%)

- 모형을 훈련시키기 위한 데이터 셋
- 이것만으로는 모형의 적합도 여부 판단 불가
- 모형에 들어가는 변수가 많아질수록 오차(ex. Mean squared error)는 줄어듦
- 지속적으로 모형을 훈련시키는 경우 오차는 지속적으로 감소
- 과적합(over-fitting)의 문제 발생
- 보통 50~60%를 선택

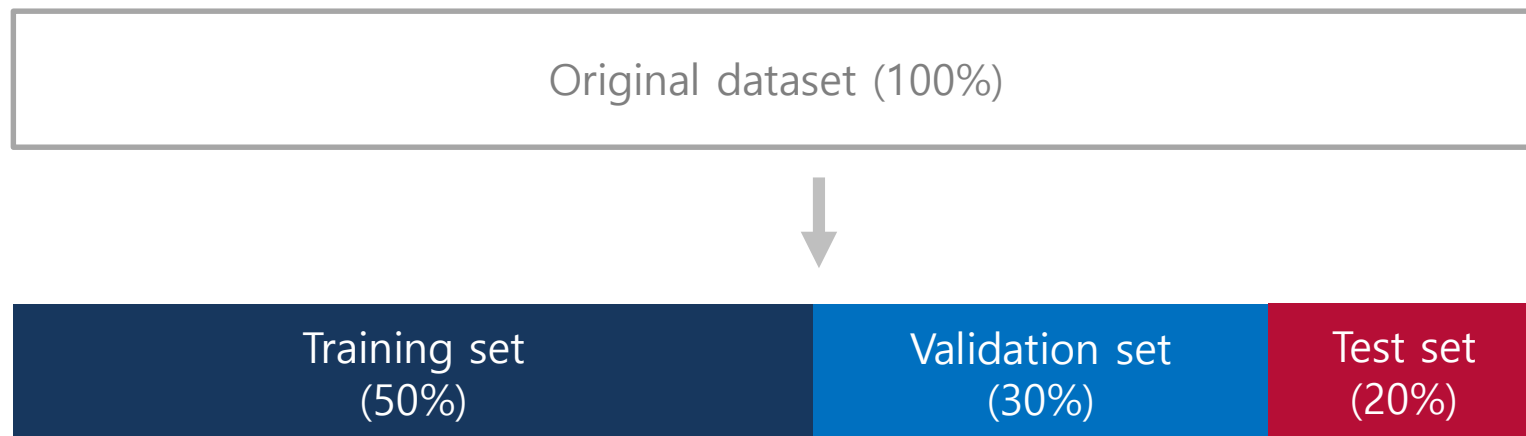
Validation set
(30%)

- Training set의 훈련모형에 validation set을 적용하여 오차 계산
- 지속적으로 적용할수록 어느 순간부터 방향을 바꾸어서 증가하게 됨
- Validation set에서의 오차 변곡점에서 stop
- 보통 20~30%로 선택

Test set
(20%)

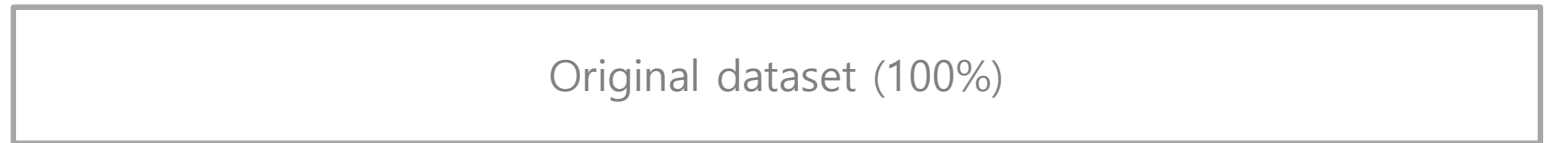
- Training&Validation set에 의해 선택된 모형에 test set을 적용하여 오차 계산
- 보통 20~30%로 선택

- 모형 평가를 위한 데이터셋 분할



출처: <http://rfriend.tistory.com>

■ 모형 평가를 위한 데이터셋 분할



Training set
(50%)

- 모형을 훈련시키기 위한 데이터 셋
- 이것만으로는 모형의 적합도 여부 판단 불가
- 모형에 들어가는 변수가 많아질수록 오차(ex. Mean squared error)는 줄어듦
- 지속적으로 모형을 훈련시키는 경우 오차는 지속적으로 감소
- 과적합(over-fitting)의 문제 발생
- 보통 50~60%를 선택

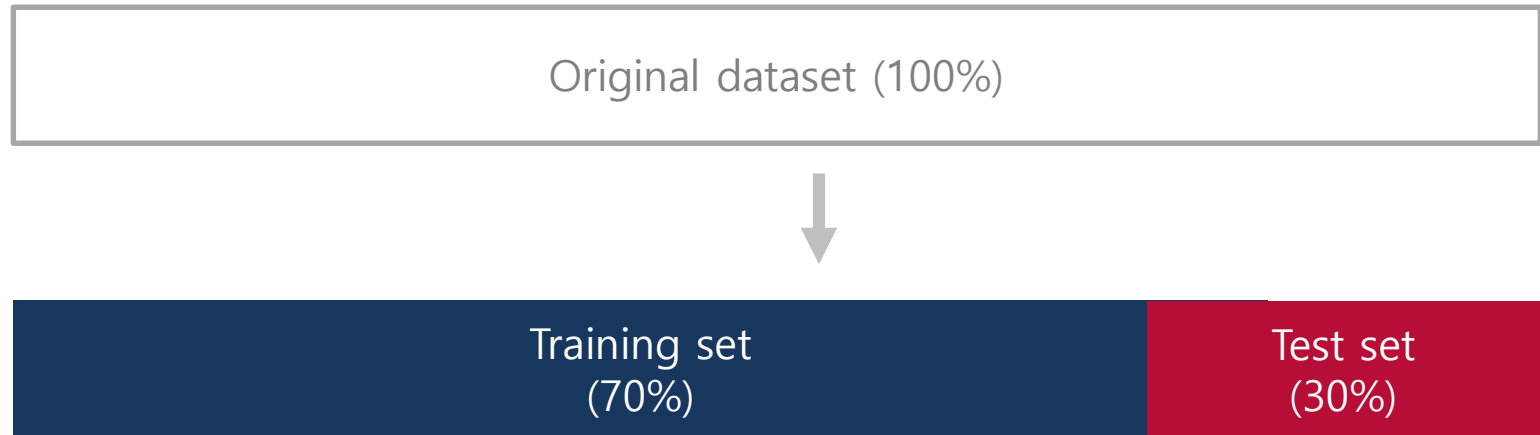
Validation set
(30%)

- Training set의 훈련모형에 validation set을 적용하여 오차 계산
- 지속적으로 적용할수록 어느 순간부터 방향을 바꾸어서 증가하게 됨
- Validation set에서의 오차 변곡점에서 stop
- 보통 20~30%로 선택

Test set
(20%)

- Training&Validation set에 의해 선택된 모형에 test set을 적용하여 오차 계산
- 보통 20~30%로 선택

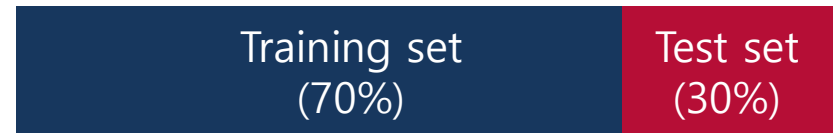
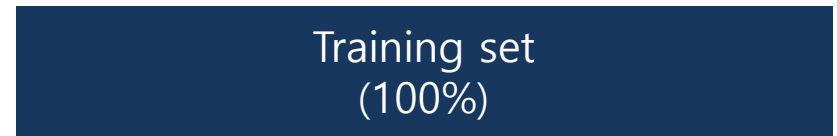
- 모형 평가를 위한 데이터셋 분할



- Rex에서 제공하는 모듈에서 적용가능한 데이터셋 형태는 크게 Training / Test로 나뉘는 것으로 간주

■ 모형 평가를 위한 데이터셋 분할

편할검증	모든 데이터를 훈련에 이용
	비율에 따라 임의로 분할
	변수로 분할



교차검증	Leave-one-out 교차 검증
	K-fold 교차검증



- 나이와 치료 방법에 따른 수술의 효과를 검정하고자 한다. 36명의 데이터를 활용하여 다중회귀분석을 수행하라. (시트명 : Op)
 - Y : 수술효과
 - age : 연령 (세)
 - treatment : 치료방법 (A, B, C)
 - split : 분할변수 (1=훈련 / 2=시험)
 - predict : 예측변수 (1=예측 / 2=훈련및 검증)

- 다음 3가지 방법에 의해 검증을 수행하시오.
 - 비율에 따라 임의로 분할 (훈련자료 70%, 시험자료 30%)
 - 변수로 분할 (변수명 split / 1=훈련, 2=시험)
 - 10-fold 교차검증

입력

Rex ▶ 회귀분석 ▶ 선형회귀분석

선형회귀분석

변수설정 | 자료분할 | 출력옵션 | 변수선택 | 그래프

데이터
전체변수
split
predict

종속변수(1개이상필수)
> Y
<

설명변수
질적변수(선택-1개이상가능)
> treatment
<

양적변수(선택-1개이상가능)
> X1
<

▼주효과 ▼교호작용

최종모형
treatment
X1
treatment:X1

삭제

☐ 상수항 포함하지 않음

도움말 | 재설정 | 확인 | 취소

선형회귀분석

변수설정 | 자료분할 | 출력옵션 | 변수선택 | 그래프

변수목록
split
predict

훈련 및 검증(필수)
☒ 분할검증
☐ 모든 데이터를 훈련에 이용
☒ 비율에 따라 임의로 분할
 훈련(train) 자료 70 %
 시험(test) 자료 30 %
☐ 변수로 분할
 분할변수(1-훈련, 2-시험)
 > <

교차검증
☐ Leave-one-out 교차검증
☒ K-fold 교차검증 K 10

예측(선택)
 분할변수(1-예측, 2-훈련 및 검증)
 > <

도움말 | 재설정 | 확인 | 취소

입력

Rex ▶ 회귀분석 ▶ 선형회귀분석

선형회귀분석

변수설정 | 자료분할 | 출력옵션 | 변수선택 | 그래프

출력

회귀계수

☒ 신뢰구간
신뢰수준 0.95

☒ 분산팽창지수(VIF)

☒ 분산분석표

제공함유형

☐ Type I ☐ Type II ☒ Type III

☐ 적합도검정
요약표

☐ 잔차진단그래프

☐ 단순회귀 추가

저장

훈련자료

☒ 적합값 ☒ 비표준화 잔차 ☒ 쿡의 거리

☒ 신뢰구간 ☒ 표준화 잔차 ☒ 헤트 행렬의 대각원소

☒ 예측구간 ☒ 스튜던트화 잔차

신뢰수준 0.95

시험자료

☐ 예측값

☐ 신뢰구간

☐ 예측구간

신뢰수준

예측자료

☐ 예측값의 신뢰구간

☐ 예측값의 예측구간

신뢰수준

☒ 자료분할지표

다음말 | 재설정

확인 | 취소

저장 > 훈련자료	설명	저장된 변수명
적합값	추정된 회귀식에 기초하여 해당 요인에 대해 예측된 종속변수 값	Fitted_train_LM
신뢰구간	주어진 관측값에 대한 적합값의 평균적인 신뢰구간	Fitted_95CI_Lower (Upper)_train_LM
예측구간	주어진 관측값에 대한 적합값이 존재할 수 있는 샘플링 오차까지 고려한 예측구간	Fitted_95PI_Lower (Upper)_train_LM
신뢰수준	신뢰구간 및 예측구간 계산 시 고려하는 신뢰수준	-
비표준화 잔차	추정된 회귀식에 의한 적합값과 실제값의 차이	unstdResid_train_LM
표준화 잔차	비표준화 잔차를 표준화한 값	stdResid_train_LM
스튜던트 잔차	해당 값을 제외한 후 계산한 표준화잔차	studResid_train_LM
쿡의 거리	해당 관측값이 제외될 때 모형의 변화정도 / 클수록 영양점 가능성 높아짐	CookDist_train_LM
헤트행렬의 대각원소	레버리지 값 (독립변수의 평균)으로부터 떨어져 있는 정도	HatValue_train_LM
자료분할지표	임의분할 시 인덱스 (Training/Testing)	Partition_idx_LM

임의분할

Validation using training/testing dataset

	N.non-missing observations	Percent	RMSE	MAE	Rsquared
Training	25	69.4444	3.4273	2.8218	0.9602
Testing	11	30.5556	4.0723	3.5924	0.9433

Validation using training/testing dataset

	N.non-missing observations	Percent	RMSE	MAE	Rsquared
Training	22	61.1111	3.3814	2.8486	0.9207
Testing	14	38.8889	4.8082	3.9371	0.9705

변수분할

10-Fold Cross Validation

- Warning : The number of parameters in model is greater than the number of samples in each fold.
The number of parameters in the model is 6

교차검증

RMSE	SD(RMSE)	Rsquared	SD(Rsquared)
4.1101	1.397	0.9447	0.0383

- 나이와 치료 방법에 따른 수술의 효과를 검정하고자 한다. 36명의 데이터를 활용하여 다중회귀분석을 수행하라. (시트명 : Op)
 - Y : 수술효과
 - X1 : 연령 (세)
 - treatment : 치료방법 (A, B, C)
 - split : 분할변수 (1=훈련 / 2=시험)
 - predict : 예측변수 (1=예측 / 2=훈련및 검증)

- split 변수를 이용하여 분할검증하고, predict 변수를 이용하여 예측값을 계산하시오.

입력

Rex ▶ 회귀분석 ▶ 선형회귀분석

선형회귀분석

변수설정 | 자료분할 | 출력옵션 | 변수선택 | 그래프

데이터

전체변수

split
predict

설명변수

종속변수(1개이상필수)
Y

질적변수(선택-1개이상가능)
treatment

양적변수(선택-1개이상가능)
X1

▼주효과 ▼교호작용

최종모델

treatment
X1
treatment:X1

삭제

☐ 상수항 포함하지 않음

도움말 | 재설정 | 확인 | 취소

선형회귀분석

변수설정 | 자료분할 | 출력옵션 | 변수선택 | 그래프

변수목록

훈련 및 검증(필수)

☒ 분할검증

☐ 모든 데이터를 훈련에 이용

☐ 비율에 따라 임의로 분할

훈련(train) 자료 %

시험(test) 자료 %

☒ 변수로 분할

분할변수(1-훈련, 2-시험)

split

☐ 교차검증

☐ Leave-one-out 교차검증

☒ K-fold 교차검증 K

예측(선택)

분할변수(1-예측, 2-훈련 및 검증)

predict

도움말 | 재설정 | 확인 | 취소

입력

Rex ▶ 회귀분석 ▶ 선형회귀분석

선형회귀분석

변수설정 | 자료분할 | 출력옵션 | 변수선택 | 그래프

출력

회귀계수

☒ 신뢰구간 ☐ 분산팽창지수(VIF)
신뢰수준

☒ 분산분석표

제곱합유형
☐ Type I ☐ Type II ☒ Type III

☐ 적합도검정 ☐ 잔차진단그래프

요약표
☐ 단순회귀 추가

저장

훈련자료

☐ 적합값 ☐ 비표준화 잔차 ☐ 쿡의 거리
☐ 신뢰구간 ☐ 표준화 잔차 ☐ 헤트 행렬의 대각원소
☐ 예측구간 ☐ 스튜던트화 잔차
신뢰수준

시험자료

☒ 예측값
☒ 신뢰구간
☒ 예측구간
신뢰수준

예측자료

☒ 예측값의 신뢰구간
☒ 예측값의 예측구간
신뢰수준

☐ 자료분할지표

도움말 | 재설정 | 확인 | 취소

저장 > 시험자료	설명	저장된 변수명
예측값	훈련자료를 통해 추정된 회귀식을 바탕으로 독립변수의 해당 관측값을 적용하여 얻은 적합값	Predicted_testing_LM
신뢰구간	예측값의 신뢰구간	Predicted_95CI_Lower(Upper)_testing_LM
예측구간	샘플링 오차를 고려한 예측값의 예측구간	Predicted_95PI_Lower(Upper)_testing_LM
신뢰수준	신뢰구간 및 예측구간 계산 시 고려하는 신뢰수준	-

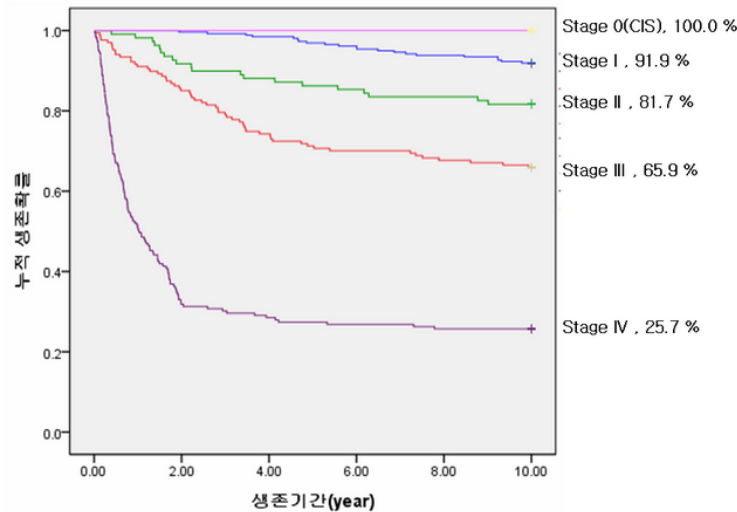
저장 > 시험자료	설명	저장된 변수명
예측값	훈련/시험자료를 통해 추정된 회귀식을 바탕으로 적용하여 얻은 예측값	Predicted_pred_LM
예측값의 신뢰구간	예측값의 신뢰구간	Predicted_95CI_Lower(Upper)_pred_LM
예측값의 예측구간	샘플링 오차를 고려한 예측값의 예측구간	Predicted_95PI_Lower(Upper)_pred_LM
신뢰수준	신뢰구간 및 예측구간 계산 시 고려하는 신뢰수준	-



노모그램의 활용

■ 노모그램 소개

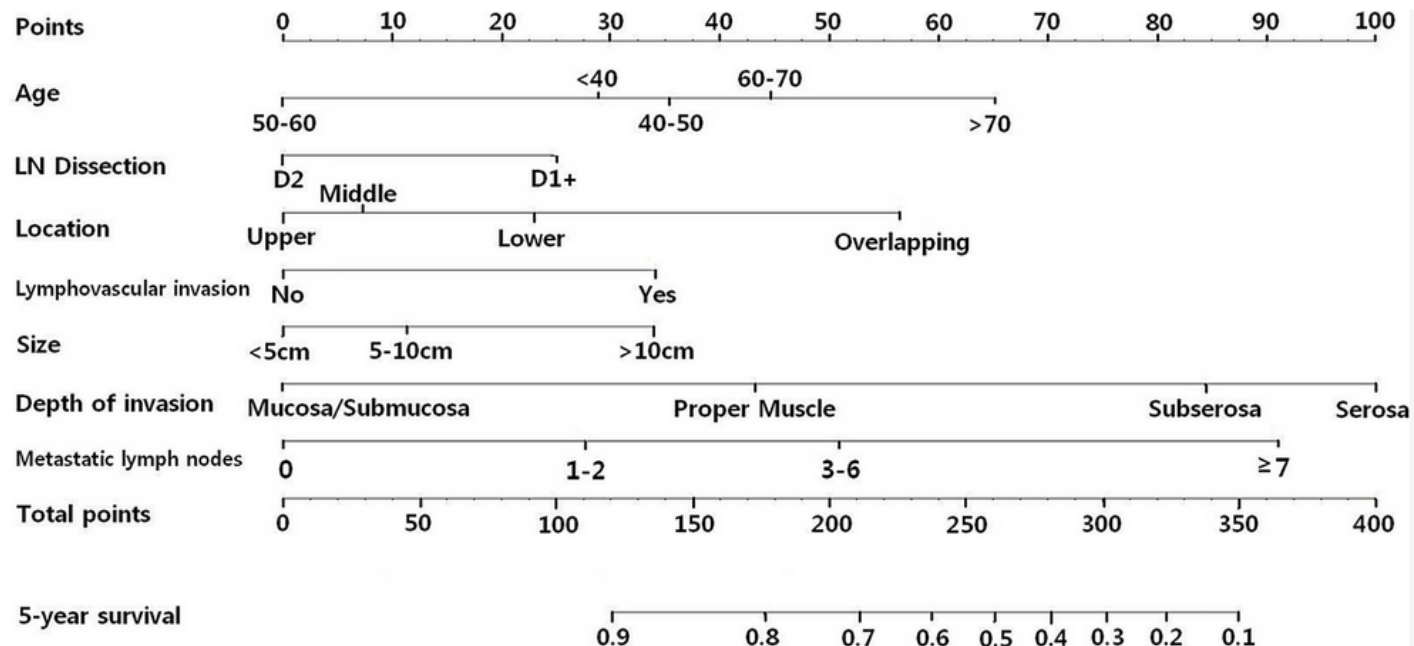
- 예측모형에 필요한 요소들과 영향력을 하나의 도표로 나타낸 것
 - ✓ 한 가지 기준만으로 병을 진단하거나 생존율을 예측하는 것은 타당한가



- 환자 1 : 위암, III기, 45세, 남자, 전신 상태 양호, 위암 절제술 받음
 - 환자 2 : 위암, III기, 55세, 여자, 전신 상태 불량 (빈혈, 당뇨 동반), 위암 절제술 받음
- 두 환자에 대하여 똑같이 10년째 생존율을 65.9%로 예측하는 것이 옳은가?
- 여러 가지 기준을 종합적으로 적용할 수 있는 도구가 필요

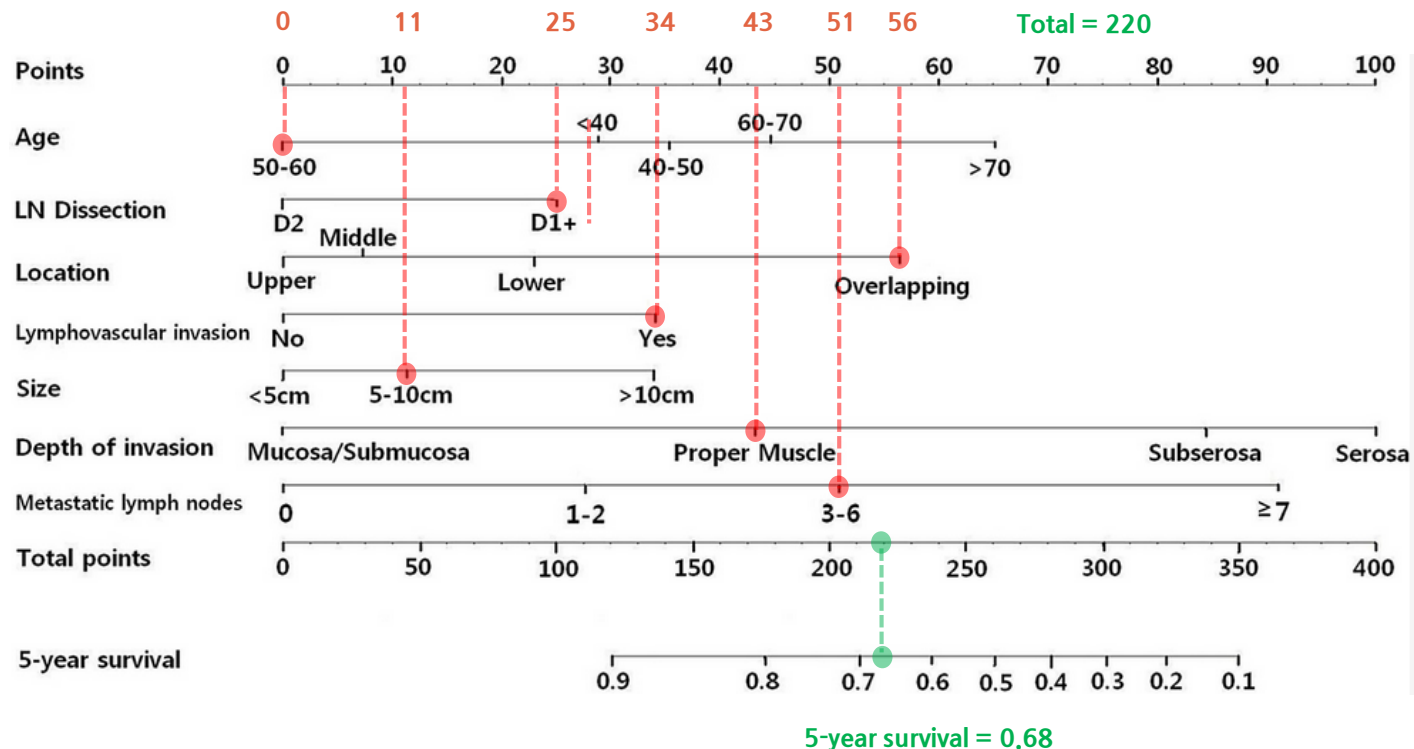
■ 노모그램 소개

- 어떤 몇 가지 설문이나 간단한 검사들의 조합을 통해 복잡하고 어려운 진단을 쉽게 내릴 수 있도록 설계된 도구
- Logistic regression model, Cox PH regression model, Poisson regression model 등을 근간으로 하여 작성



■ 노모그램 소개

- 어떤 몇 가지 설문이나 간단한 검사들의 조합을 통해 복잡하고 어려운 진단을 쉽게 내릴 수 있도록 설계된 도구
- Logistic regression model, Cox PH regression model, Poisson regression model 등을 근간으로 하여 작성



Pulmonary Hypertension: A Nomogram Based on CT Pulmonary Angiographic Data for Prediction in Patients without Pulmonary Embolism¹

Galit Aviram, MD
Hezzy Shmueli, MD
Sharon Z. Adam, MD
Achiude Bendet, MD
Tomer Ziv-Baran, PhD
Arie Steinvil, MD
Abraham Shlomo Berliner, MD, PhD
Nachum Neshet, MD
Yanai Ben-Gal, MD
Yan Topilsky, MD

Purpose:

To use cardiovascular data from computerized tomographic (CT) pulmonary angiography for facilitating the identification of pulmonary hypertension (PH) in patients without acute pulmonary embolism.

Materials and Methods:

The institutional human research committee approved this retrospective study; informed consent was waived. Patients without pulmonary embolism who underwent CT pulmonary angiography and echocardiography within 24 hours of each other between December 2008 and October 2012 were retrospectively identified. The diameters of the

Figure 1

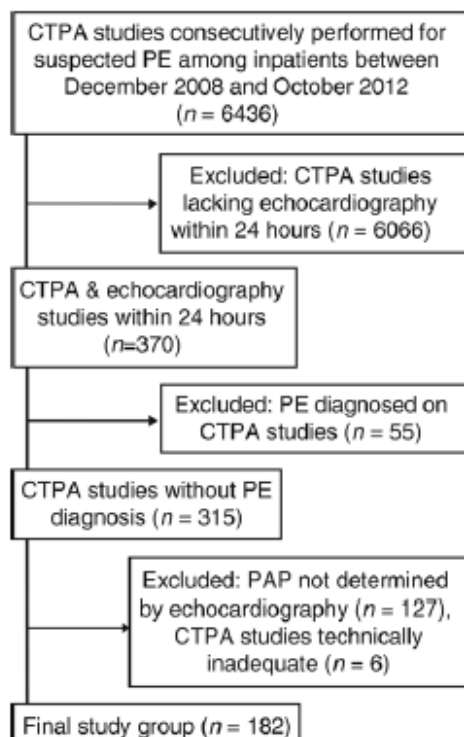


Figure 1: Flowchart shows patient selection. CTPA = CT pulmonary angiography, PAP = PA pressure.

Table 1

Patient Characteristics

Characteristic	Study Group (n = 182)	Test Group (n = 82)
Age (y)*	68.5 ± 18.8	69.5 ± 19.3
Male sex	73 (40.1)	26 (31.7)
Obesity	20 (11.0)	8 (9.7)
Diabetes mellitus	42 (23.1)	23 (28.0)
Hypertension	102 (56.0)	50 (61.0)
Hyperlipidemia	73 (40.1)	31 (37.8)
Chronic obstructive pulmonary disease	45 (24.7)	9 (11.0)
Asthma	9 (4.9)	6 (7.3)
Other lung	13 (7.1)	1 (1.2)
Smoker	78 (42.9)	24 (29.3)
Congestive heart failure	37 (20.3)	19 (23.2)
Chronic renal failure	18 (9.9)	7 (8.5)
Ischemic heart disease	45 (24.7)	15 (18.3)
Cerebrovascular accident or transient ischemic attack	25 (13.7)	7 (8.5)
Hospitalization within past 3 months	41 (22.5)	26 (31.7)
Active malignancy	20 (11.0)	12 (14.6)
Past malignancy	19 (10.4)	8 (9.7)
Active deep venous thrombosis	5 (2.7)	0 (0)
Past deep venous thrombosis	7 (3.8)	5 (6.1)
Reason for CT referral		
Dyspnea	143 (78.6)	63 (76.8)
Chest pain	27 (14.8)	9 (11.0)
Cough	3 (1.6)	1 (1.2)
Other	9 (4.9)	9 (11.0)

Note.—Unless otherwise indicated, data are number of patients, with percentages in parentheses.

* Data are means ± standard deviation.

Table 2

CT Measurements according to Presence of Pulmonary Hypertension

Parameter	All (n = 182)	PH (n = 98)	No PH (n = 84)	PValue
Age (y)	75.0 (59.0–82.0)	79.0 (70.0–84.3)	66.0 (40.5–80.0)	<.001
Male sex*	73 (40.1)	37 (37.8)	36 (42.9)	.484
Reflux grade*				
1	69 (37.9)	28 (28.6)	41 (48.8)	.004
2	41 (22.5)	18 (18.4)	23 (27.4)	
3	23 (12.6)	16 (16.3)	7 (8.3)	
4	27 (14.8)	19 (19.4)	8 (9.5)	
5	12 (6.6)	10 (10.2)	2 (2.4)	
6	10 (5.5)	7 (7.1)	3 (3.6)	
Right ventricular volume (cm ³)	113 (91–147)	122 (94–161)	108 (86–136)	.027
Right ventricular short diameter (mm)	43 (38–48)	45 (40–51)	40 (36–45)	<.001
Right atrial volume (cm ³)	93 (73–120)	110 (81–139)	79.0 (70–101)	<.001
Left ventricular volume (cm ³)	75 (58–98)	72 (57–102)	77 (60–96)	.829
Left ventricular short diameter (mm)	44 (40–49)	44 (40–50)	44 (38–48)	.194
Left atrial volume (cm ³)	87 (65–112)	99 (75–123)	73 (58–99)	<.001
Main PA diameter (mm)	29 (26–32)	31 (28–33)	27 (24–30)	<.001
PA-to-aorta ratio	0.87 (0.79–0.97)	0.90 (0.81–0.99)	0.84 (0.76–0.95)	.012
Right atrial to left atrial volume ratio	1.08 (0.92–1.29)	1.07 (0.90–1.34)	1.08 (0.95–1.28)	.834
Right ventricular to left ventricular volume ratio	1.08 (0.92–1.29)	1.67 (1.39–2.01)	1.48 (1.25–1.68)	.001

Note.—Unless otherwise indicated, data are medians, with interquartile ranges in parentheses.

* Data are number of patients, with percentage in parentheses.

Table 3

The Distribution of Age and CT Measurements and Accuracy for Discrimination

Variable	All (n = 182)	PH (n = 98)	No PH (n = 84)	P Value	Area Under the Curve*
Age (y)					
<67	64 (35)	19 (19)	45 (54)	<.001	0.67 (0.59, 0.75)
≥67	118 (65)	79 (81)	39 (46)		
Reflux grade					
<3	110 (60)	46 (47)	64 (76)	<.001	0.65 (0.57, 0.77)
≥3	72 (40)	52 (53)	20 (24)		
Right ventricular volume (cm ³)					
<142	129 (71)	59 (60)	70 (83)	.001	0.62 (0.53, 0.70)
≥142	53 (29)	39 (40)	14 (17)		
Right ventricular short diameter (mm)					
≤43	92 (51)	36 (37)	56 (67)	<.001	0.65 (0.57, 0.73)
≥43	90 (49)	62 (63)	28 (33)		
Right atrial volume (cm ³)					
<106	115 (63)	43 (44)	72 (86)	<.001	0.71 (0.63, 0.79)
≥106	67 (37)	55 (56)	12 (14)		
Left atrial volume (cm ³)					
≤71	59 (32)	19 (19)	40 (48)	<.001	0.70 (0.62, 0.77)
72–108	69 (8)	37 (38)	32 (38)		
>108	54 (30)	42 (43)	12 (14)		
Main PA diameter (mm)					
<28	75 (41)	22 (22)	53 (63)	<.001	0.71 (0.63, 0.79)
≥28	107 (59)	76 (78)	31 (37)		
PA-to-aorta ratio					
<0.86	84 (46)	33 (34)	51 (61)	<.001	0.64 (0.56, 0.72)
≥0.86	98 (54)	65 (66)	33 (39)		

Note.—Unless otherwise indicated, data are number of patients, with percentage in parentheses.

* Data in parentheses are 95% confidence intervals (CIs).

ciated with PH. The probability of having PH could be calculated by using the following equation:

$$PPH = 1/(1 + e^z),$$

where *PPH* is the probability of having PH, *e* is 2.718, and

$$z = -2.556 + 1.495 \cdot A + 0.926 \cdot PA + 0.776 \cdot PA/AO + 0.968 \cdot RF + 1.278 \cdot RA,$$

where *A* is the value for patient age (age ≥ 67: *A* = 1; age < 67: *A* = 0), *PA* is the pulmonary artery diameter value (diameter ≥ 28 mm: *PA* = 1; diameter < 28: *PA* = 0), *PA/AO* is the ratio of the *PA* diameter to that of the aorta (ratio ≥ 0.86: *PA/AO* = 1; ratio < 0.86: *PA/AO* = 0), *RF* is the reflux value (reflux ≥ 3: *RF* = 1; reflux < 3: *RF* = 0), and *RA* is the right atrial volume value (right atrial volume ≥ 106: *RA* = 1; right atrial volume < 106, *RA* = 0).

P = .285). Receiver operating characteristic analysis allowed identification that a threshold value of 0.4 provided sensitivity of 90%. Cross-validation (10-fold) showed sensitivity of 85.7%, specificity of 60.7%, positive predictive value of 71.3%, negative predictive value of 76.1%, and accuracy of 73.1%, which were sufficient for use of this model as a screening tool for PH in patients who undergo CT pulmonary angiography. Figure 3 illustrates the distribution of probabilities produced with the model for the presence of PH in patients with and without evidence of PH at echocardiography.

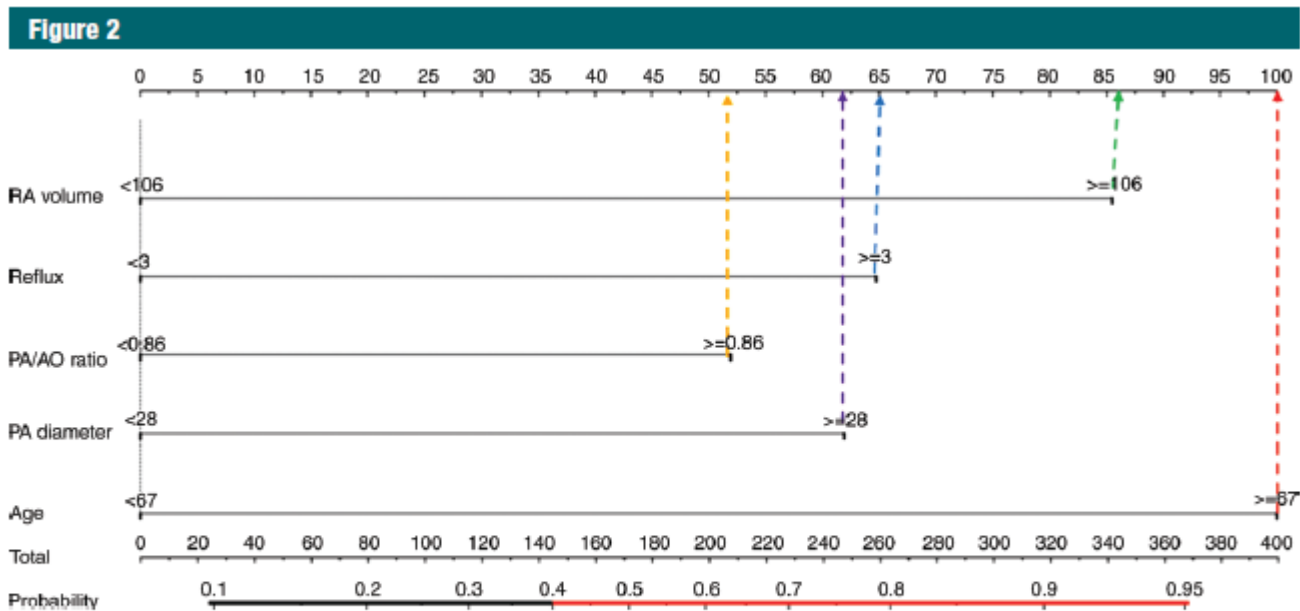


Figure 2: Nomogram shows logistic model for prediction of PH. Each CT-derived measurement has corresponding value (points) that appear in upper toolbar (ie, right atrial [RA] volume > 106 = 86 points (green line), reflux grade ≥ 3 = 65 points (blue line), PA diameter > 28 mm = 62 points (purple line), and PA-to-aorta ratio > 0.86 = 52 points (yellow line), in addition to patient's age (> 67 = 100 points, red line). Summarized total was applied on bottom scale to obtain probability of PH. Any probability greater than 0.4 was compatible with PH. AO = aorta.

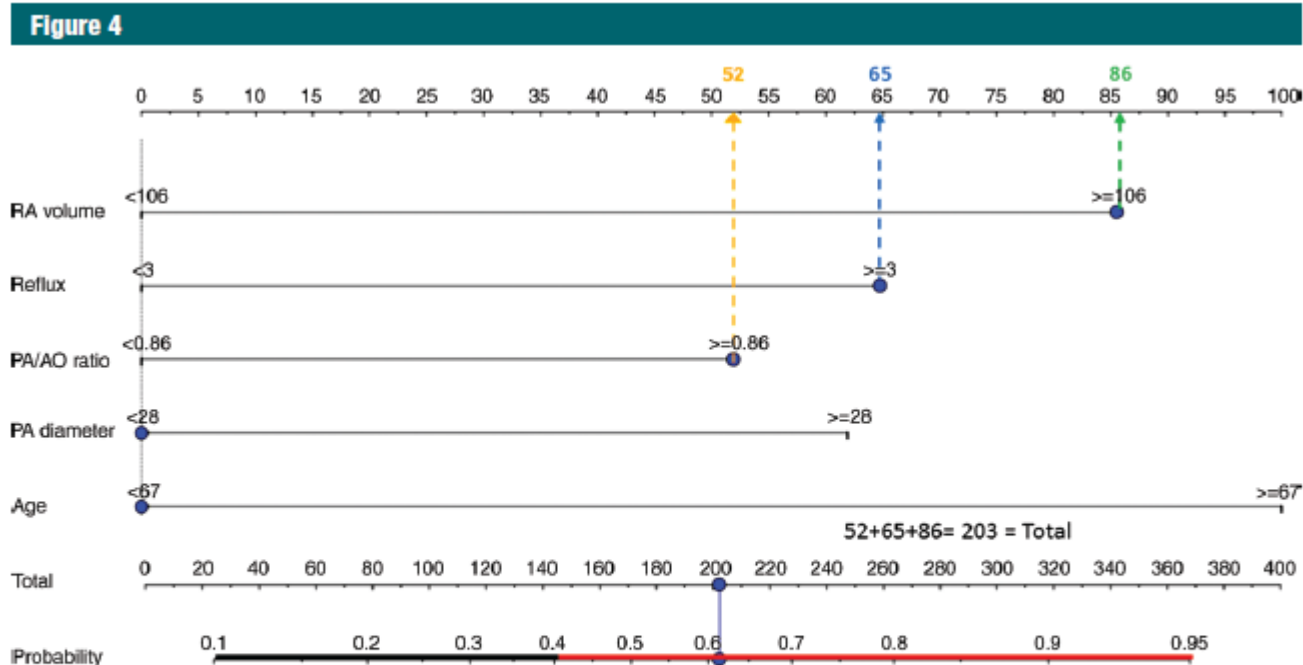
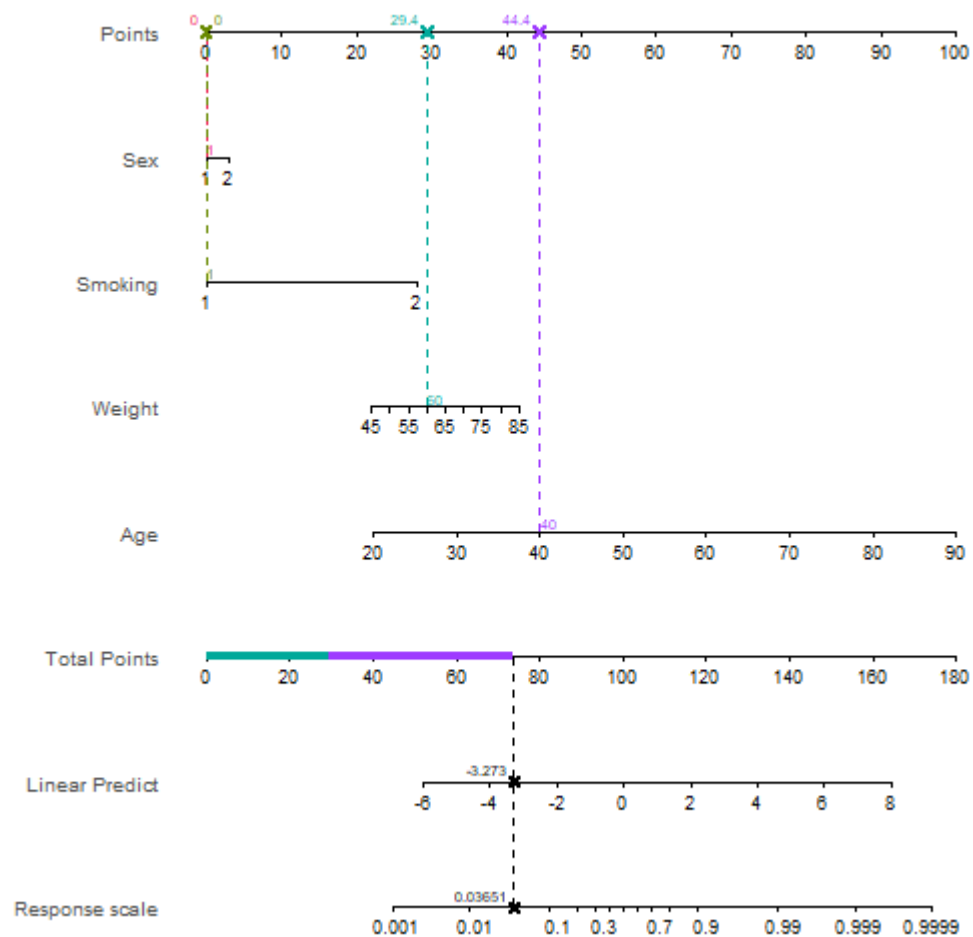


Figure 4: Nomogram for a 30-year-old woman with negative results for PE and PA systolic pressure reading of 62 mm Hg at echocardiography. Adding the corresponding points of the parameters (age, 30 years = 0 points; right atrial [RA] volume, 150 mL = 86 points [green line]; reflux grade, 4 = 65 points [blue line]; PA diameter, 27.8 mm = 0 points; PA to aorta [AO] ratio, 0.87 = 52 points [yellow line]) yields a total of 203 points. According to nomogram, her probability of having PH is 0.61. Because probability of greater than 0.4 was defined as being compatible with PH, nomogram allowed correct prediction of presence of increased PA pressure.

- 심장병 센터를 방문한 성인 33명에게서 관상동맥질환 발병 여부와 연령, 성별, 체중, 흡연력을 조사하였다. 관상동맥질환의 위험요인을 분석하고, 4가지 요인을 모두 포함하는 노모그램을 작성하시오. 이후 비흡연인 40세 남성의 체중이 60kg일 때, 이 사람의 관상동맥질환 발병확률을 노모그램 위에 나타내시오.

- CHD = 관상동맥질환 유무 (0=없음 / 1=있음)
- Age = 연령 (year)
- Sex = 성별 (1=남자 / 2=여자)
- Weight = 체중 (kg)
- Smoking = 흡연력 (1=비흡연 / 2=흡연)



Rexsoft

감사합니다

| 문의



<http://rexsoft.org>



help@rexsoft.org

홈페이지 '질문과 답변' 게시판을 통해 Rex 설치, 다운로드, 기능 등 사용문의를 남겨주세요.
신속하고 친절할 상담을 통해 사용자들의 궁금증에 답변해 드립니다.



Make Analysis Easy and Fast