Forum Norr Statistikernätverk

Lunchseminarium 11/3 2024

From Data to Decision: A Guide to Developing and Assessing Risk Prediction Models

Wendy Yi-Ying Wu

wendy.wu@regionvasterbotten.se

Registercentrum Norr

Kliniska Studier Sverige Forum Norr

Outlines

- What is the risk prediction model?
- How and when to use it?
- How to develop a risk prediction model?
- How to evaluate?

Kliniska Studier Sverige Forum Norr

Predictor finding studies (risk factor or prognostic factor studies)

Aim to identify which predictors independently contribute to the prediction of a diagnostic or prognostic outcome

Prediction model studies

Aim to develop, validate or update a multivariable prediction model

Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019 Jan;170(1):51–8.

Kliniska Studier Sverige Forum Norr

Risk prediction models

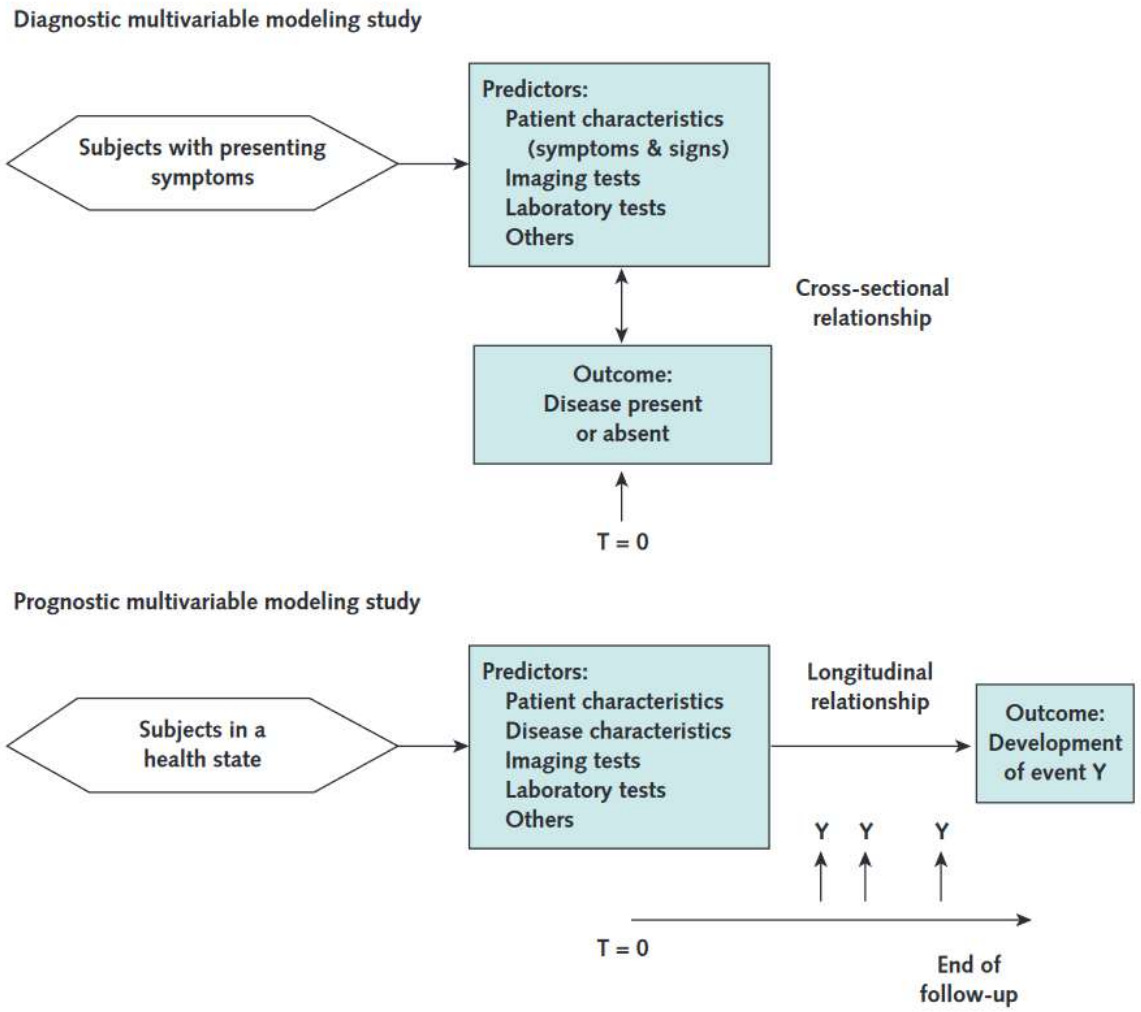
- Use predictors (covariates) to estimate the **absolute probability or risk** in an individual with a particular predictor profile

- Predictors
  - subject characteristics (eg, age and sex), examination results, imaging, electrophysiology, blood, urine, genetic markers, proteins and metabolites, etc.

Kliniska Studier Sverige Forum Norr

TRIPOD statement

Box A. Schematic representation of diagnostic and prognostic prediction modeling studies.



Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1–73.

Kliniska Studier Sverige Forum Norr

Purpose

- to guide healthcare professionals and individuals in their decision making
  - decide whether we need further testing by predicting the probability of the underlying disease
  - decide whether we need to start a treatment/use more intensive treatment/delay a treatment
  - decide whether we need a surgery by balancing short-term risks (e.g., 30-day mortality) and long-term risks (e.g., long-term survival risk)

- to inform individuals about their risks of having (diagnosis) or developing (prognosis) a particular disease or outcome.

PMID:32175364

Kliniska Studier Sverige Forum Norr

Examples

Journal of
Clinical Medicine

MDPI

Article

Loss of Smell and Taste Can Accurately Predict COVID-19 Infection: A Machine-Learning Approach

$$x = -1.76 + 0.88 \times ((1 \text{ if VAS for loss of smell} \geq 21) \text{ or } (0 \text{ if VAS for loss of smell} < 21)) + 1.83 \times ((1 \text{ if VAS for loss of taste} \geq 44) \text{ or } (0 \text{ if VAS for loss of taste} < 44)) + 0.79 \times ((1 \text{ if VAS for dyspnea} \geq 28) \text{ or } (0 \text{ if VAS for dyspnea} < 28)) + 0.61 \times ((1 \text{ if fever}) \text{ or } (0 \text{ if no fever})) + 0.70 \times ((1 \text{ if diarrhea}) \text{ or } (0 \text{ if no diarrhea})) - 1.13 \times ((1 \text{ if female}) \text{ or } (0 \text{ if male}))$$

Kliniska Studier
Sverige
Forum Norr

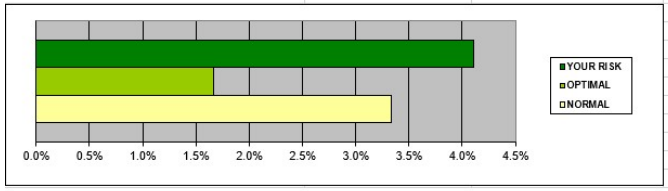
Examples

- Framingham risk score

From The Framingham Heart Study — **Enter Values Here**

General CVD Risk Prediction

Risk Factor	Units	(Type Over Placeholder Values in Each Cell)	Notes
Sex	male (m) or female (f)	f	
Age	years	45	
Systolic Blood Pressure	mmHg	135.0	
Treatment for Hypertension	yes (y) or no (n)	n	
Smoking	yes (y) or no (n)	n	
Diabetes	yes (y) or no (n)	n	
HDL	mg/dL	45	
Total Cholesterol	mg/dL	180	
Your 10-Year Risk (The risk score shown is derived on the basis of an equation. Other print products, use a point-based system to calculate a risk score that approximates the equation-based one.)		4.1%	*If value is < the minimum for the field, enter the minimum value. If value is > the maximum for the field, enter the maximum value.*
Your Heart/Vascular Age		49	



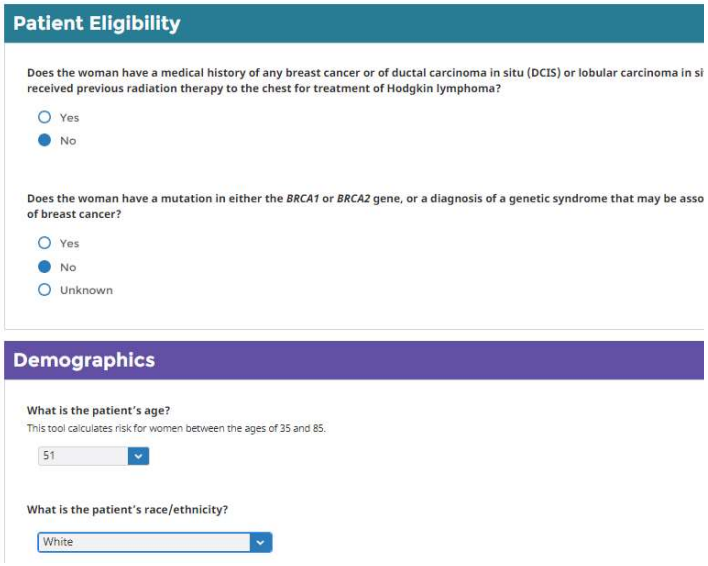
Calculator prepared by R.B. D'Agostino and M.J. Pencina based on a publication by D'Agostino et al. in Circulation

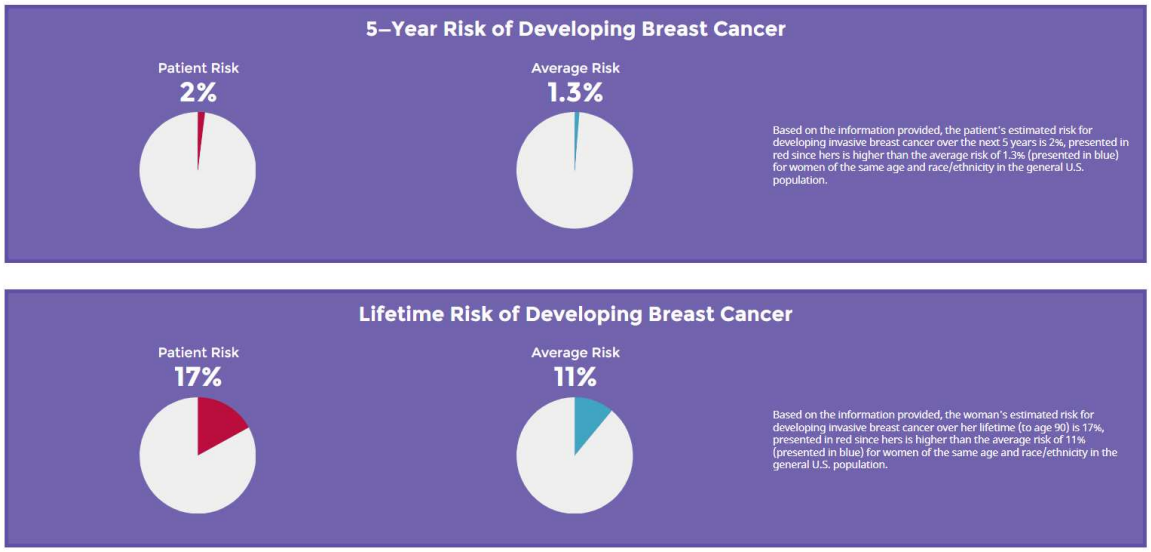
https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/

Kliniska Studier Sverige Forum Norr

Examples

- Breast cancer risk assessment model (The Gail Model)



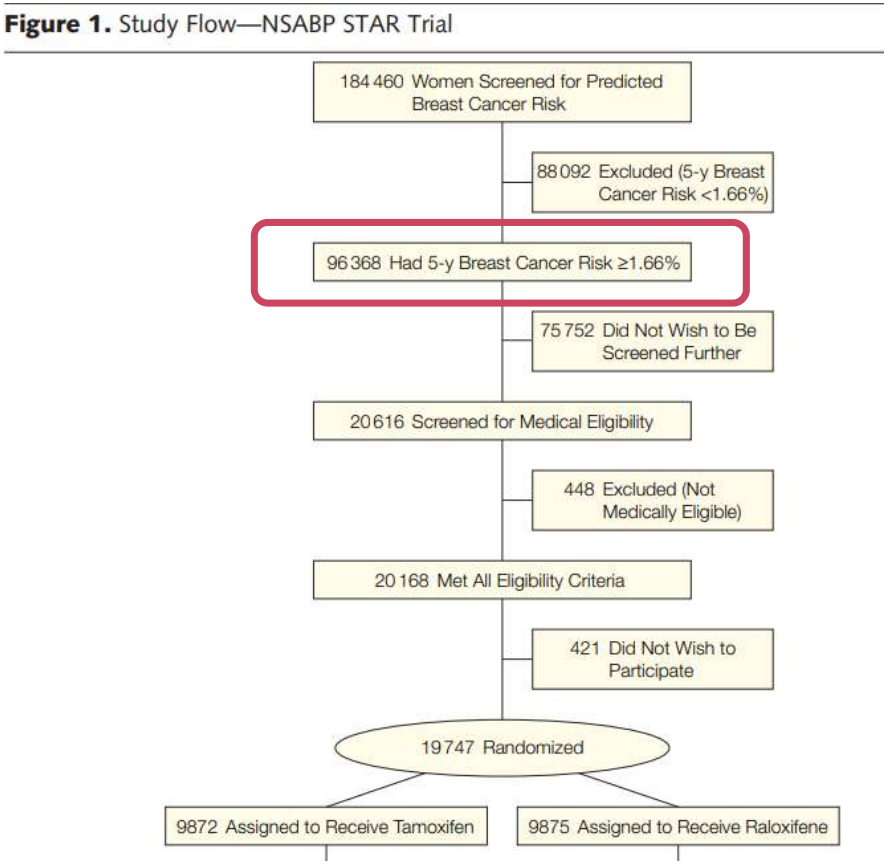


https://bcrisktool.cancer.gov/calculator.html

Kliniska Studier Sverige Forum Norr

Applications

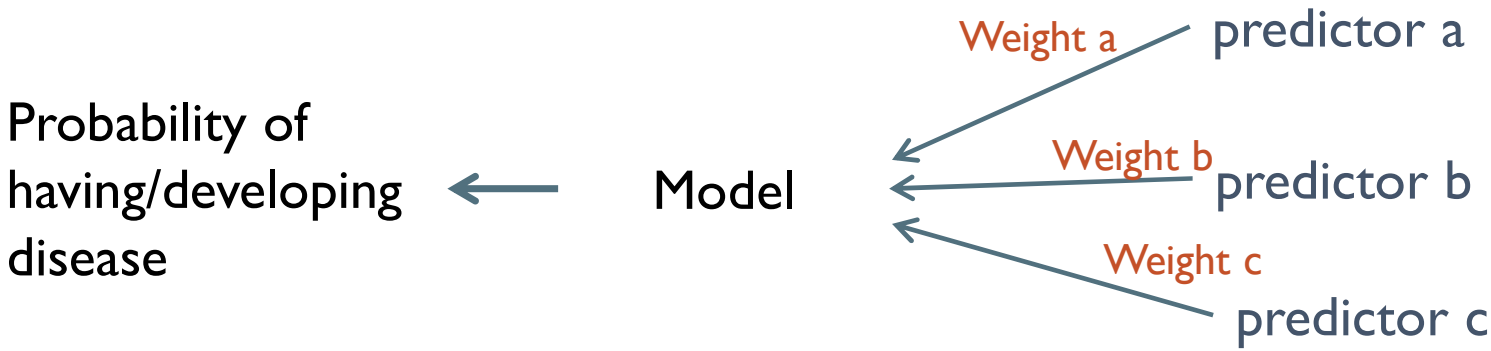
- Select appropriate participants for randomized controlled trials



PMID: 16754727

Kliniska Studier Sverige Forum Norr

Risk prediction model



Kliniska Studier Sverige Forum Norr

Developing a risk prediction model

- Identification of the important predictors (out of a set of preselected candidate predictors)
- Assigning the relative weights for each predictor in a combined risk score
- Estimating the model's predictive performance
  - Including its calibration, discrimination and reclassification properties
- Assessing its potential for optimism using so-called internal validation techniques
- Adjusting the model for over fitting if necessary

Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012 May 1;98(9):683–90.

Kliniska Studier Sverige Forum Norr

Source of data

- Cohort, (nested case-control, case-cohort)
- Registry data
- RCT

- Case-control study (only for diagnostic multivariable model)

Kliniska Studier Sverige Forum Norr

Candidate predictors

- Theoretically, all variables suspected of being associated with the outcome of interest could be considered as candidate predictors, but this association does not need to be causal.
- Existing knowledge of previously established predictors
- Clearly defined and measured in a standardized and reproducible way
  - Lower measurement error or inter-observer variability (reliability, consistency)
- More pragmatic (applicability, availability and cost)
  - Develop a model that is applicable in daily practice→ use predictors that are in line with daily practice
  - Quite readily available, not too costly to obtain, and can be measured with reasonable precision

Kliniska Studier Sverige Forum Norr

Analysis

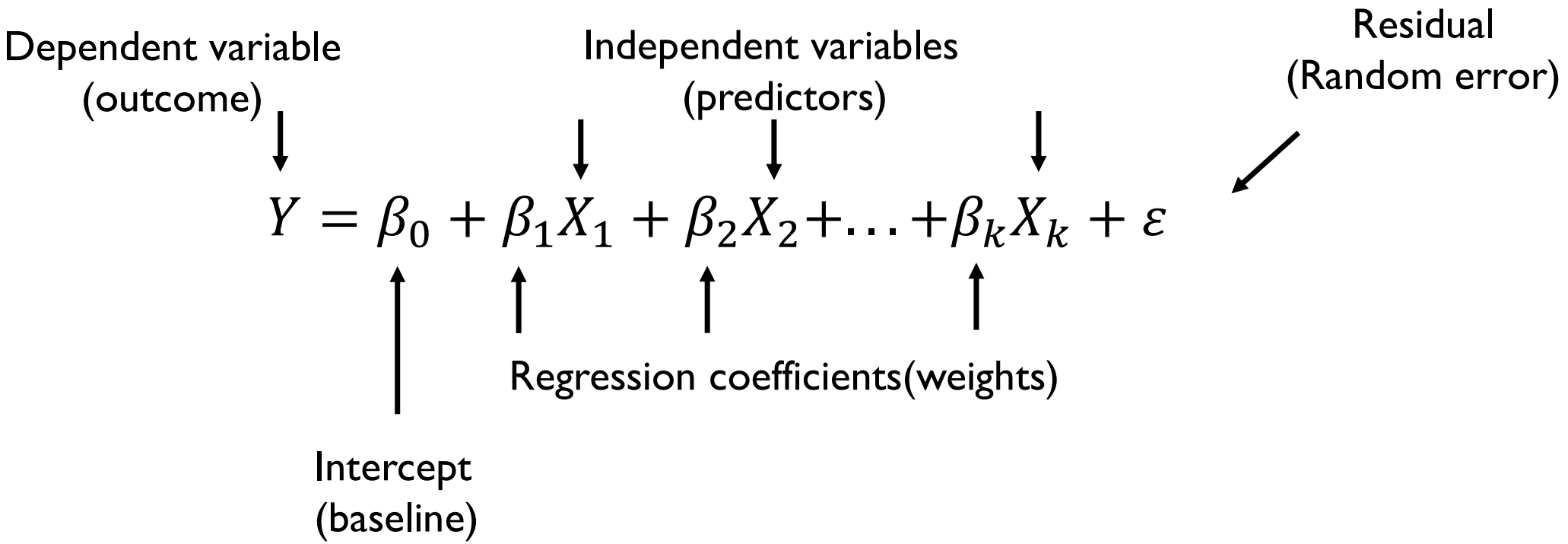
- Missing data:
  - Imputation of missing values often yields less biased results
- Continuous predictors
  - Should not be turned into dichotomies and linearity should not be assumed
- Model

Spline

Type of response variable	Model
Continuous	Linear regression
Binary	Logistic regression
Time to event	Cox PHs regression

Kliniska Studier Sverige Forum Norr

Linear regression



Kliniska Studier Sverige Forum Norr

Logistic regression

Dependent variable (outcome)

Independent variables (predictors)

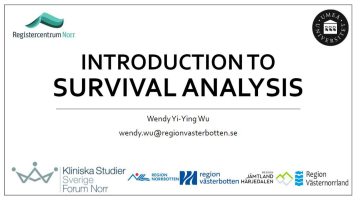
$$log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

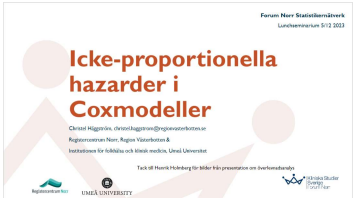
Regression coefficients (weights)

Intercept (baseline)

Kliniska Studier Sverige Forum Norr

Cox proportional hazards regression





Dependent variable (outcome)

Independent variables (predictors)

$$h(t) = h_0(t) \times \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

baseline hazard

Regression coefficients (weights)

Kliniska Studier Sverige Forum Norr

Analysis

- Too many predictors but too little data (p>>n)
  - Problems?
    - False positive findings
    - Poor validation
  - Solution:
    - Sample size calculation
    - EPV (event per variable): 1 to 10 (20) rule of thumb is often applied.
    - Data reduction: variable clustering
      - Combined similar predictors to a single one: all different types of CVD history
      - Principle component analysis/ hierarchical clustering

Type of response variable	Model	Number of predictors
Continuous	Linear regression	Total sample size / 15
Binary	Logistic regression	$\text{Min}(n_1,n_2)$ /15
Time to event	Cox PHs regression	Number of failures / 15

Harrell, F. E., Jr. (2016). *Regression modeling strategies*. Springer International Publishing.

Developing the final model

- There is no consensus about the best method of arriving at the final model.
  - Full model
    - all candidate predictors are included in the final prediction model
  - Predictor selection strategy
    - Backward elimination better than forward selection

Risk scores

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

$$log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$h(t) = h_0(t) \times \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

Kliniska Studier Sverige Forum Norr

Model performance

Discrimination

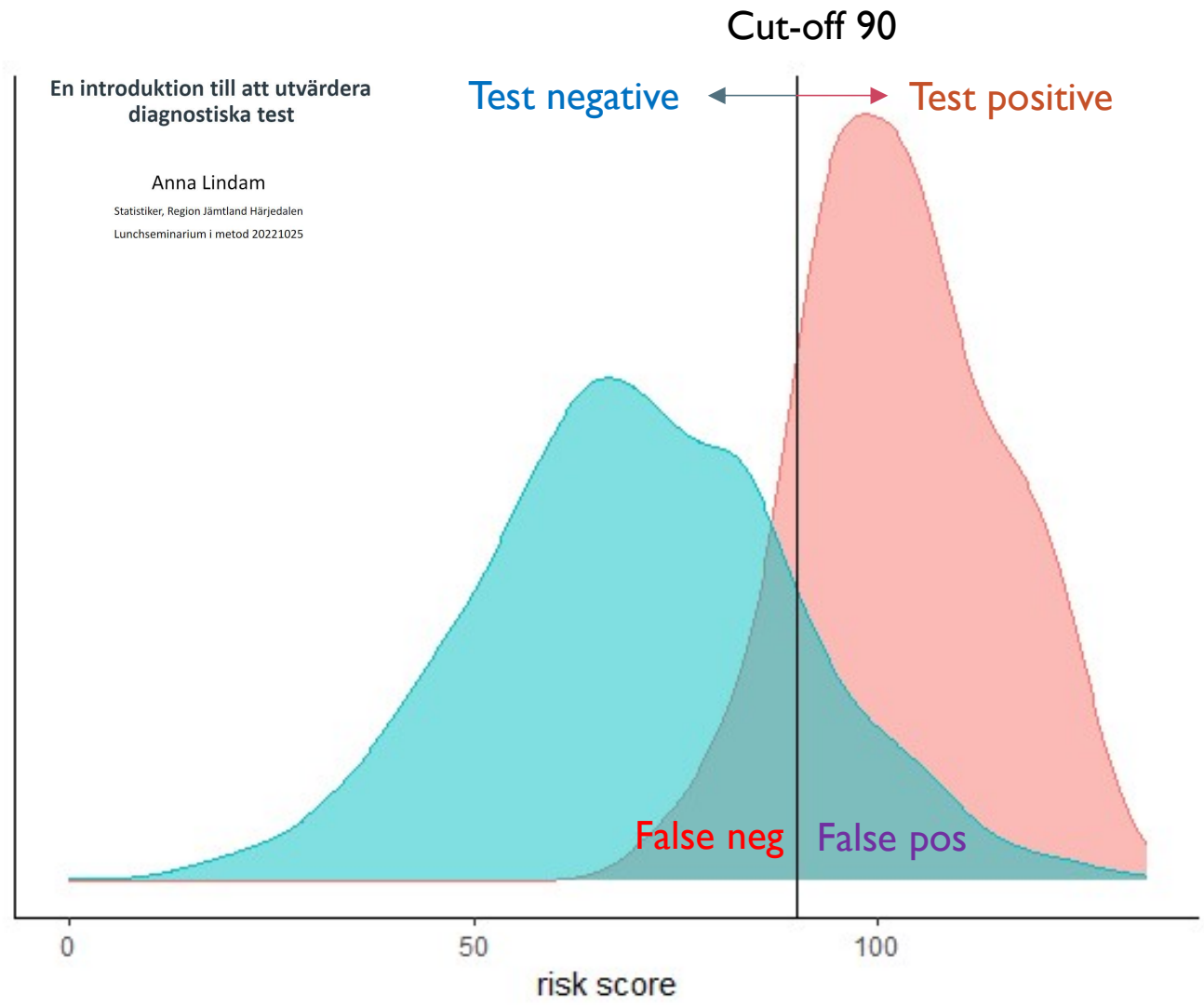
- How well the model differentiates individuals who experienced the outcome from those who remained event free

Calibration

- Agreement between prediction probability and observed outcome frequencies

Kliniska Studier Sverige Forum Norr

Discrimination



	Disease status	
Test (90)	Yes	No
Positive	a	c
Negative	b	d

Sensitivity: a/(a+b)
Specificity: d/(c+d)

case

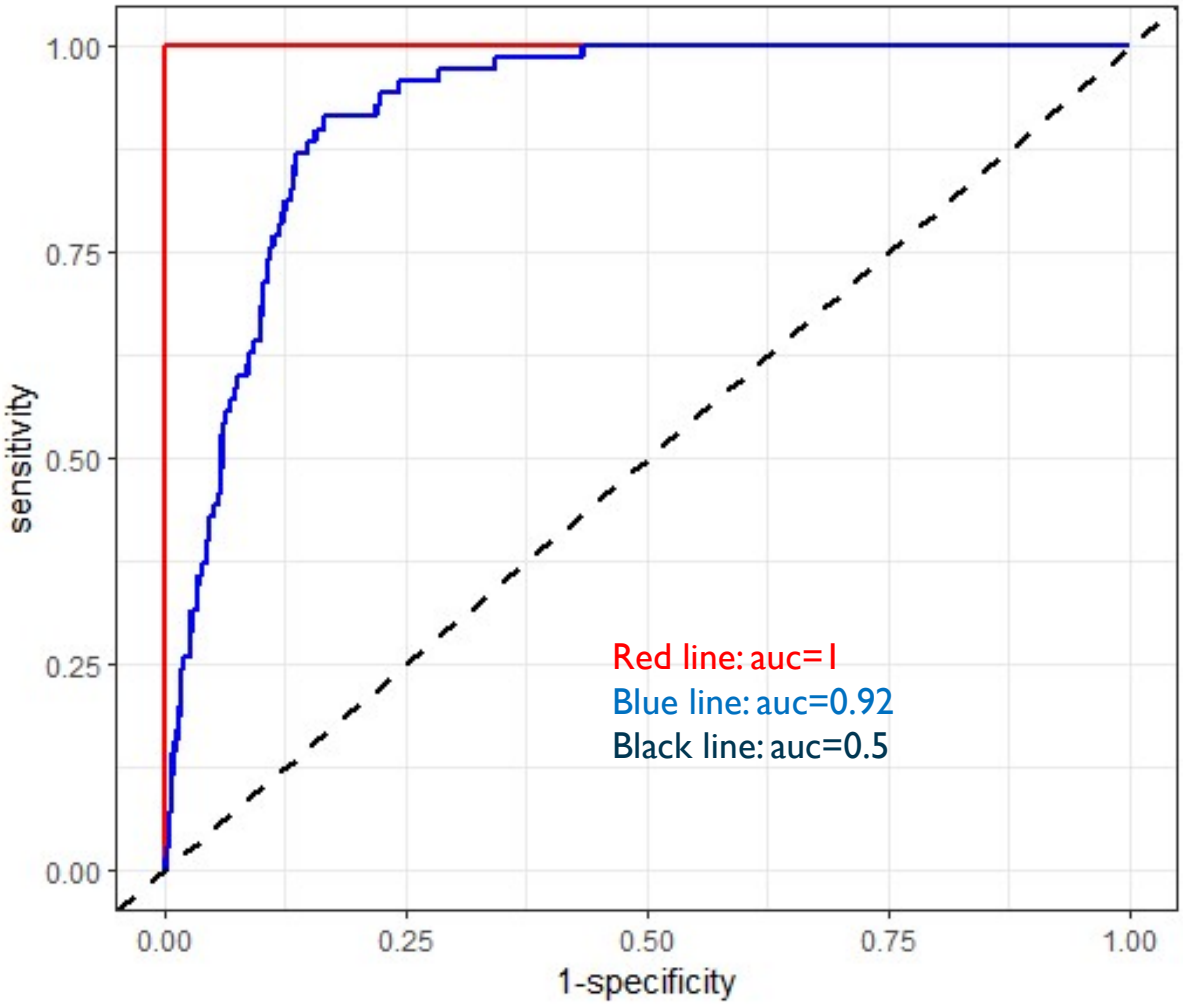
	Disease status	
Test (80)	Yes	No
Positive	a+x	c+y
Negative	b-x	d-y

Sensitivity: (a+x)/(a+b)
Specificity: (d-y)/(c+d)

Sverige
Forum Norr

Discrimination

Receiver operating characteristic (ROC)



Area under curve (c-statistics):

- the chance that given two individuals, one who will develop the event of interest and one who will remain event free, the prediction model will assign a higher probability of an event to the former.

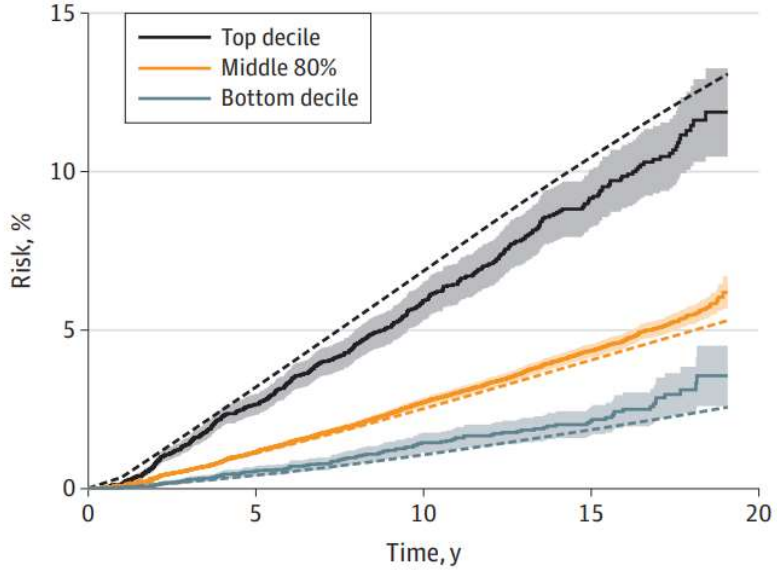
- C statistics
  - <0.6 poor discrimination
  - 0.6-0.75, possibly helpful discrimination
  - >0.75 clearly helpful discrimination

Kliniska Studier Sverige Forum Norr

Calibration

the accuracy of absolute risk estimates

A Tyrer-Cuzick model



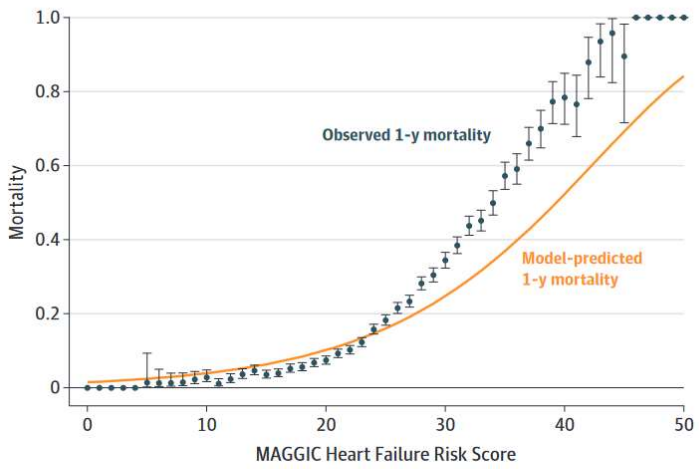
No. at risk

Top decile	13214	6331	3234	1545
Middle 80%	105712	55209	30492	16232
Bottom decile	13213	6616	3504	1766

Solid lines indicate observed risk;
broken lines indicate expected risk

PMID: 29621362

Figure 2. Observed and Predicted 1-Year Mortality



The graph represents the relationship between observed (the data markers represent the mean and the error bars represent the 95% CI) and predicted 1-year mortality using the Meta-analysis Global Group in Chronic Heart Failure (MAGGIC) risk score (orange line). The model underestimated mortality in patients with a predicted mortality greater than 30%. Adapted from Sartipy et al.[29]

PMID: 29049590

Kliniska Studier Sverige Forum Norr

Internal Validation

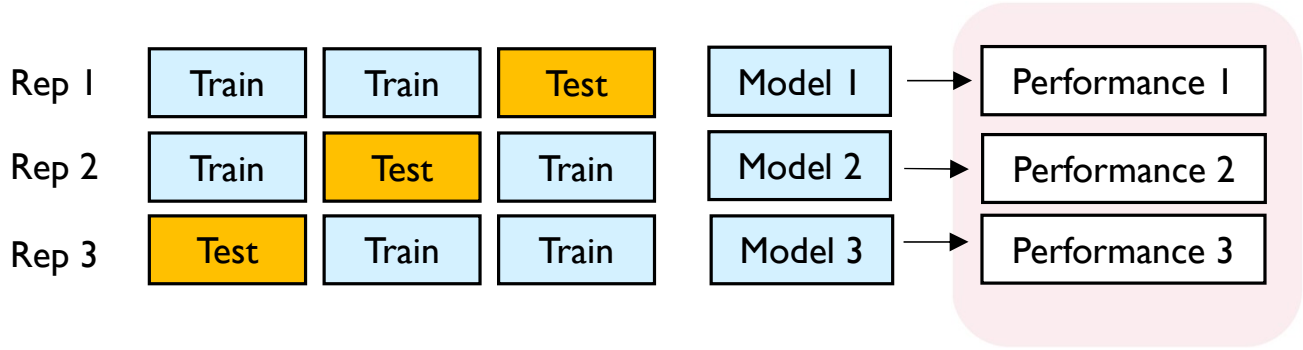
- The model was designed to optimally fit the development samples and it becomes less accurate when tested in new but similar individuals (overfitting)
  → Models might yield optimistic apparent performance
  - Optimism in model performance increases when the number of predictors increases and the number of events decreases
- Internal validation
  - to estimate the potential for overfitting and optimism in model performance
  - to calculate a developed prediction model's reproducibility for the derivative sample and protects current data from being misinterpreted.

Kliniska Studier Sverige Forum Norr

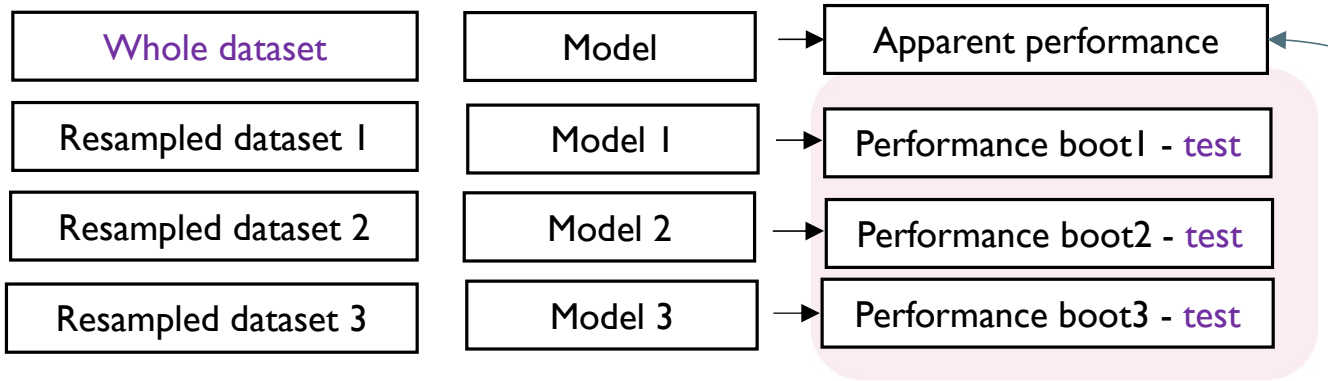
Internal Validation

- Split-sample
  - 2/3 trained dataset and 1/3 validation
- Cross-validation

Statistically inefficient because not all data are used to produce the prediction model



- Bootstrapping



Things to consider

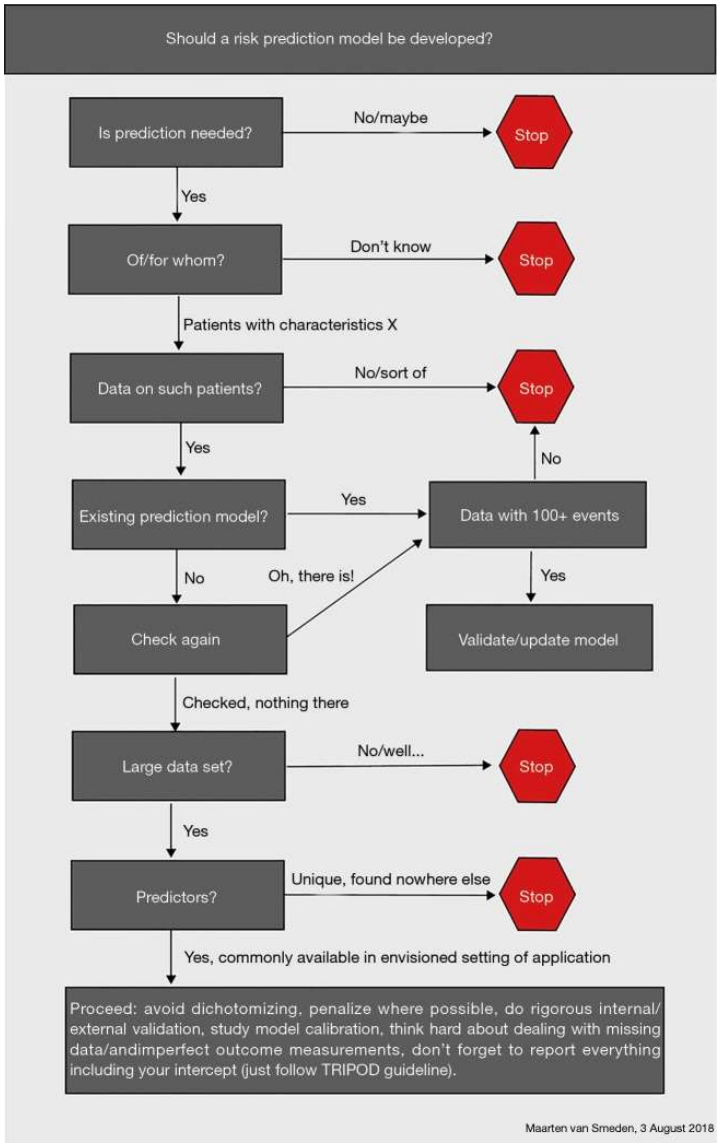
Don't do this

- Predictors: what I have in hand
- Analysis: excluding the participants with missing values
- Analysis: categorised a continuous variable
- Validation: randomly splitting a single data set into model development and model validation data sets

Do this

- Predictors: formulating good hypotheses that lead to specification of relevant candidate predictors and possible interactions.
- Analysis: imputing the missing values if possible.
- Analysis: using a continuous variable and allow nonlinearity
- Validation: using entire sample for model development and bootstrapping for internal validation

Kliniska Studier Sverige Forum Norr

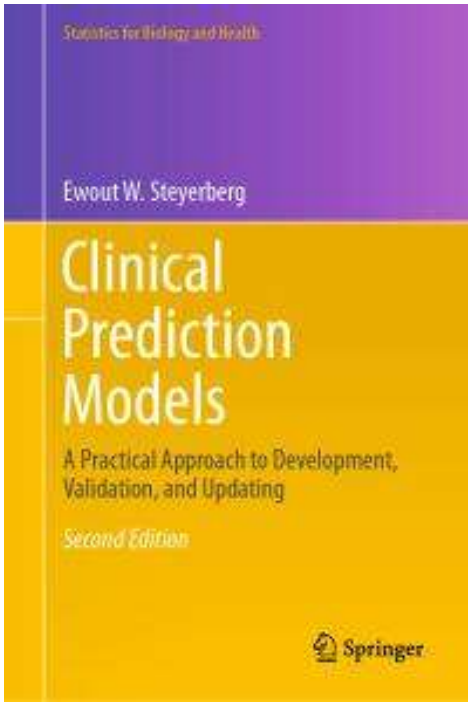


Should a risk prediction model be developed?

https://twitter.com/MaartenvSmeden/status/1025315100796899328

Kliniska Studier Sverige Forum Norr

Suggested references



- Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012 May 1;98(9):683–90.
- Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012 May 1;98(9):691–8.
- Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1–73.
- Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019 Jan;170(1):51–8.

Kliniska Studier Sverige Forum Norr

Coming Seminars

- Canvas page: https://www.canvas.umu.se/courses/2600
- **10/4 kl 12-13 Rensning och samkörning av forskningsdata**

  Christel Häggström, Region Västerbotten/ Institutionen för folkhälsa och klinisk medicin, Umeå universitet
- **7/5 kl 12-13 Metoder för upprepade mätningar**

  Staffan Betnér, Region Västerbotten/ Institutionen för folkhälsa och klinisk medicin, Umeå universitet
- **30/5 kl 12-13 Korstabeller och andra jämförelse över andelar**

  Anna Lindam, Region Jämtland Härjedalen / Institutionen för folkhälsa och klinisk medicin, Umeå universitet

Kliniska Studier Sverige Forum Norr

Kliniska Studier Sverige – Forum Norr

Utveckling och stöd för kliniska studier i hälso- och sjukvården

Kliniska Studier Sverige Forum Norr