Development of Framingham cardiovascular risk prediction algorithms

Michael J. Pencina, PhD
Boston University, Biostatistics
NHLBI's Framingham Heart Study
Harvard Clinical Research Institute

Context

- Risk prediction algorithm:
  - Input: levels of risk factors (age, sex, blood pressure etc.)
  - Output:
    - Probability of adverse event (i.e. cardiovascular disease) in a given horizon in the future (i.e. 1 year, 10 years, 30 years etc.)
    - Heart/vascular age – you have a heart of an 'x'-year-old with "standard" cardiovascular risk profile

Context

- First algorithms developed for cardiovascular disease 40 years ago based on Framingham Heart Study data
- Designed to aid but not replace clinicians
- Developed for a broad variety of conditions, including cardiovascular disease and its sub-components, diabetes, CVD risk factors and many others

Framingham Heart Study

- One of the longest-running observational studies
- Three cohorts first recruited in the early 1950s, 1970s and 2000s
- Total of over 15,000 participants with 5-60 years of follow-up
- Examined every 2-8 years
- Pioneer of cardiovascular risk prediction

Risk Assessment Considerations

- Population of Interest
- Definition of outcome
- Duration of follow-up: time horizon
- Mathematical model
- Risk factors
- Model Performance Metrics
- Validation and Transportability
- Presentation of Results
- Search for New Risk Markers

Population of Interest

- Participants free of CVD (CHD, stroke, IC, CHF)
- Participants free of CHD
- Free of cancer or diabetes?

- For diabetes risk prediction:
  - Free of diabetes?
  - Fasting glucose < 100mg/dL?
  - Other exclusions?

Definition of outcome

- In cardiovascular field multiple outcome definitions due to changes in practice and focus:
  - Full CVD (CHD, stroke, IC, CHF, CV death)
  - Hard CVD (MI, stroke, CV death)
  - CHD
  - Hard CHD (MI, CV death)
- What do clinicians and patients really care about?

Duration of follow-up

- 10-year risk a standard in CVD risk prediction
- But for recurring event shorter duration might be needed
- For young people, especially women, 10-year horizon shows very little risk
- 30-year or lifetime risk better?
- “Long-term” models which update risk factors regularly are in fact short-term models

Mathematical Model

- Cox proportional hazards regression is the current favorite
- Logistic regression useful when follow-up is short
- Parametric models (i.e. Weibull) preferred for health economists due to cost modeling opportunities
- Tree-based methods usually inferior

Mathematical Model

- I encourage limiting the number of interaction terms as they do not validate well
- Use more conservative p-value for interactions to rectify this problem
- Test only interactions of interest, with underlying reasons
- Cannot interpret main effects

Mathematical Model

- Increased interest in discovering correct shape of relationship
- Reluctance to non-linear terms
- Impact of quadratic terms harder to interpret
- Categorizing continuous variables preferred but inefficient
- Log-transformation helps reduce undue influence of extreme observations

Risk Factors

- Include standard, established risk factors before considering novel ones
- Ease and accuracy of measurement needs to be taken into account
- For CVD we include SBP, treatment, total and HDL cholesterol, diabetes, smoking
- Sex-specific or sex-pooled?
- Age as risk factors or scale?

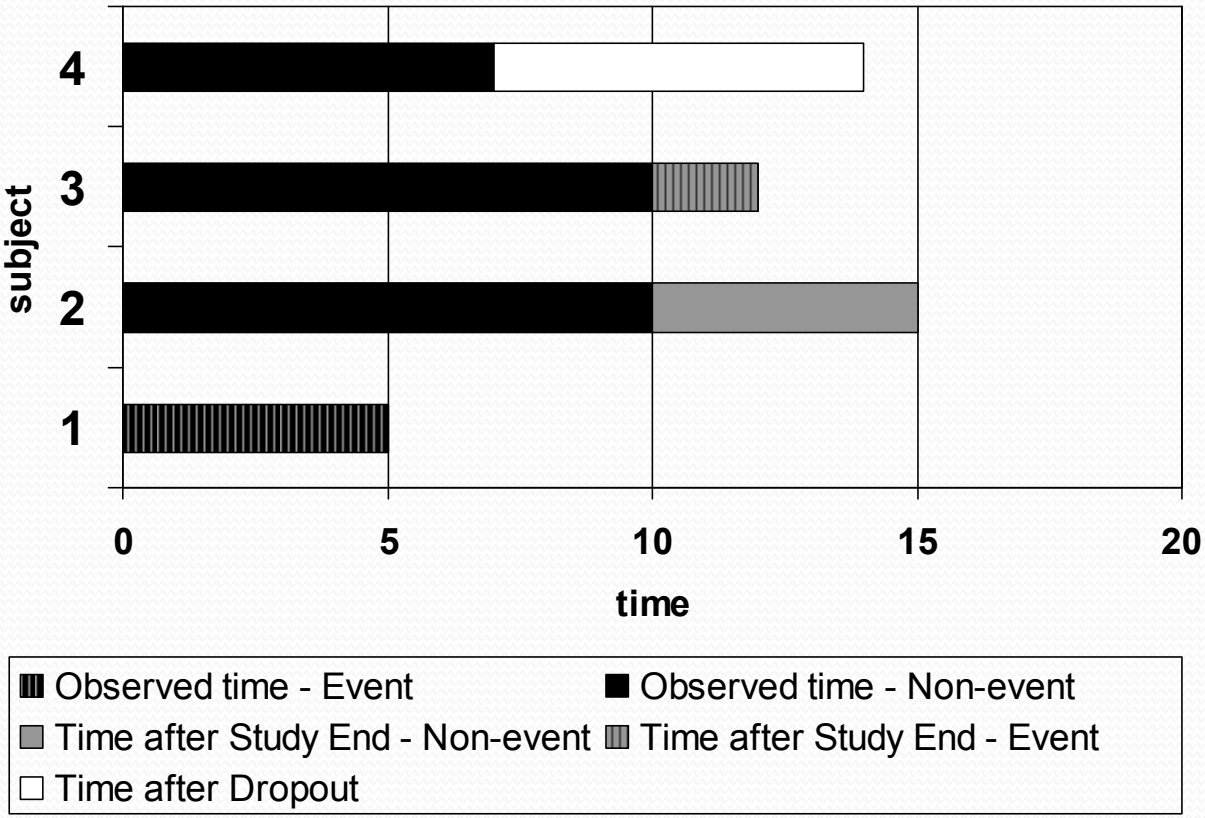
Performance Metrics

- Statistical significance of all predictors and the model is necessary
- It does not tell us much about performance
- Hazards ratio very popular as this is what Cox produces
- Odds ratios too often confused with relative risks
- Absolute more important than relative?

Performance Metrics

- Discrimination: ability of model to distinguish events from non-events;
- In longer-term survival definition extended as ability to classify people according to observed event times based on predicted probabilities
- 'C statistic' often used as measure of discrimination of risk prediction models

Different survival experience



AUC definition for binary outcomes

- $C = P(Z_i > Z_j \mid D_i=1, D_j=0)$,

where:

Z_i, Z_j are model-based risks (i.e., linear predictors)

D_i, D_j are event indicators for two subjects;

- Note that only event vs. non-event comparisons are made

Ignoring time-to-event

- Simplest extension of AUC to survival data ignores time-to-event;
- It treats censored individuals and/or drop-outs as non-events

Harrell's* C

- Any two subjects are comparable if:

  $T_i > T_j$ or $T_i < T_j$

  where T denotes survival time
- Any two subjects are concordant if:

  $T_i > T_j$ and $Z_i < Z_j$ or $T_i < T_j$ and $Z_i > Z_j$
- C statistic defined as probability of concordance given comparability

*Harrell et al., StatistMed 1996, Pencina and D'Agostino, StatistMed 2004

Discrimination

- Other measures of discrimination are gaining popularity
- Discrimination slope is a current favorite:
  - Check how far average of predictions for those we subsequent events is from the average for nonevents
  - The higher, the better
  - Values depend on proportion of events

Calibration

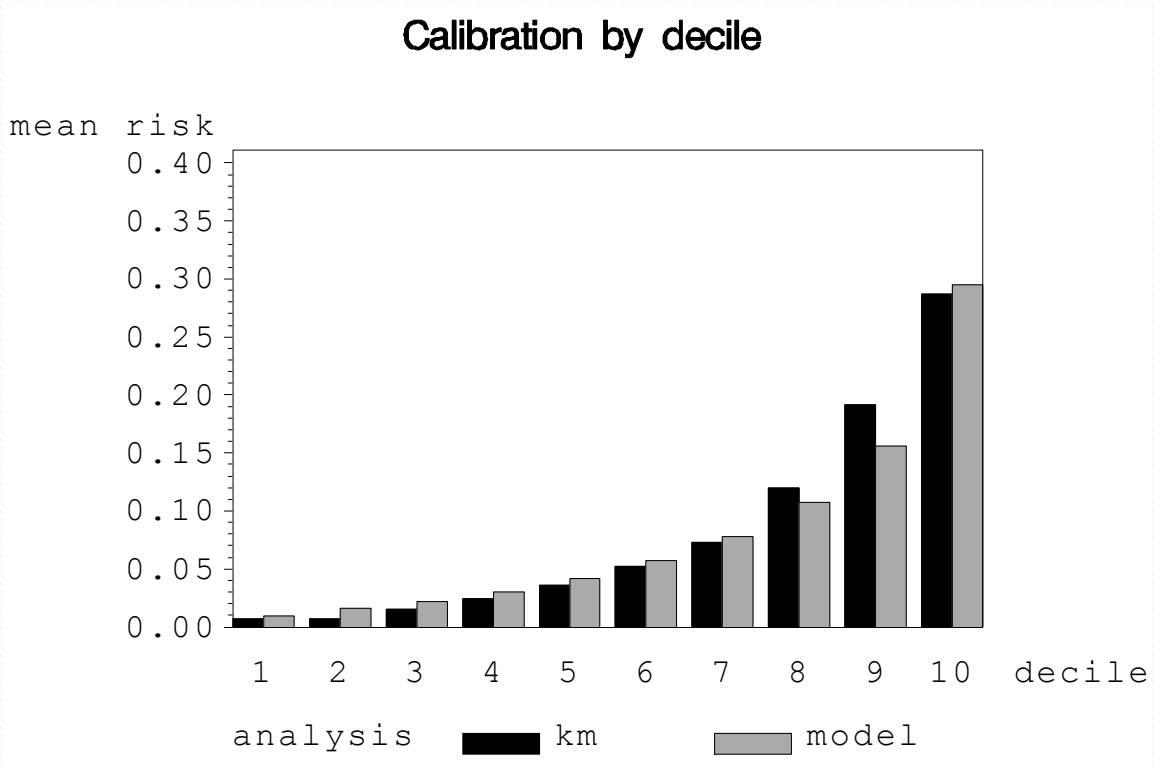
- If risk prediction is of primary interest correct calibration is essential
- Different degrees of calibration:
  - Calibration at large: are the means of predictions equal to incidence rate? – most basic, if this one fails, all other ones fail
  - Calibration by decile – Nam and D'Agostino* (extended Hosmer-Lemeshow's idea to survival
  - Linear Over-dispersion

*Nam and D'Agostino, Handbook of Statistics, 2004

Calibration

- If we had a very large sample, we could calculate 10-year risks for every age (many of them, each for different combination of risk factors)
- We would like their average for a given age to be close to the true 10-year event rate for people of this age
- For smaller sample, we use deciles of risk

Framingham Example: cvd in women chi-square=9.9 p=0.35



Validation

- How well would my algorithm do if not evaluated on the same sample on which it was developed?
- Ideally algorithm validated on external sample from the same population (i.e. Framingham algorithm on ARIC or CHS data*)
- Cross-validation and bootstrap re-sampling are good options for "internal-validation"
- 2:1 sample split often performed but inferior

*D'Agostino, JAMA 2001

Transportability

- Will my algorithm perform well if applied to a different ethnic group or geographical region?
- For example, Framingham functions developed on Caucasian cohort were applied to:
  - African Americans
  - Different European cohorts
  - Chinese population

Transportability

- Transporting to different populations might require re-calibration
- Simple re-calibration multiplies predicted probabilities by a constant
- Re-discrimination generally not possible
- Framingham functions (sometimes recalibrated) did very well when transported to different cohorts*

*D'Agostino, JAMA 2001, Liu, JAMA 2004, Marrugat, J. Epidemiol Community Health 2003

Presentation of Results

- Type of presentation:
  - Relative risk for each risk factor
  - 10-year absolute risk for risk factor combinations
  - Heart/vascular age
- Method of presentation:
  - Formula in the manuscript
  - Approximate point system
  - MS excel or other application calculator

Example 1: 10-year general CVD

- 10-year risk of general cardiovascular disease*
- Population of interest: people free of broadly defined CVD (cardiovascular death, MI, angina, stroke, TIA, vascular disease, heart failure), age 30-74
- Definition of outcome: general CVD as above
- Time horizon: 10 years
- Mathematical model: Cox regression

*D'Agostino et al., Circulation 2008

Example 1: 10-year general CVD

- Separate functions for women and men
- Risk factors:
  - Age, SBP, BP treatment, Total and HDL cholesterol, smoking and diabetes status
- Performance metrics:
  - C statistic (very good: 0.79 for men, 0.76 for men)
  - Calibration by decile (very good: 7.8 for women, 13.5 for men, both < 20)

*D'Agostino et al., Circulation 2008

Example 1: 10-year general CVD

- Presentation of results
  - Approximate points-based score for 10-year risk
  - Approximate points-based score for heart/vascular age
  - Exact MS Excel calculator for 10-year risk and heart/vascular age
  - Available at: http://www.framinghamheartstudy.org/risk/gencardio.html
- Validation: so far only internal with bootstrap

*D'Agostino et al., Circulation 2008

Example 2: 30-year hard CVD

- 30-year risk of hard cardiovascular disease*
- Population of interest: people free of broadly defined CVD (as before) and cancer, age 20-59
- Definition of outcome: hard CVD (cardiovascular death, MI and stroke)
- Time horizon: 30 years
- Mathematical model: modified Cox regression accounting for competing risk of death

*Pencina et al., Circulation 2009

Example 2: 30-year hard CVD

- One function for women and men, no interactions
- Risk factors:
  - Age, sex, SBP, BP treatment, Total and HDL cholesterol, smoking and diabetes status
- Performance metrics:
  - C statistic (very good: 0.80)
  - Calibration by decile (very good: 4.2)

*Pencina et al., Circulation 2009

Example 2: 30-year hard CVD

- Presentation of results
  - Exact MS Excel calculator for 30-year risk
  - Soon to be available on Framingham website
- Validation: internal with cross-validation and 2:1 split; external validation in progress
- Combining 10-year risk functions does not lead to accurate estimation of 30-year risk

*Pencina et al., Circulation 2009

Example 3: 4-year risk of diabetes

- 4-year risk of incident diabetes based on 28 years of follow-up*
- Population of interest: people free diabetes (fasting glucose above 126 or diabetest treatment), age 18-70
- Definition of outcome: incident diabetes
- Time horizon: 4 years

*Meigs et al., NEJM 2008

Example 3: 4-year risk of diabetes

- Mathematical model: pooled logistic regression for correlated data
- One function for women and men, no interactions
- Risk factors:
  - Age, sex, family history of diabetes, BMI, fasting glucose, SBP, HDL cholesterol, triglycerides, genotype score

*Meigs et al., NEJM 2008

Example 3: 30-year hard CVD

- Performance metrics:
  - C statistic (very good: 0.90)
  - Calibration by decile (very good: 1.9)
- Presentation of results
  - Table of relative risks
- Validation: none yet

*Meigs et al., NEJM 2008

Search for new markers

- In the last two decades researchers identified numerous new candidate risk markers and postulated their inclusion into the risk score algorithms
- There is no agreement, however, how to measure the added utility of these new markers beyond what is offered by the standard risk factors

Statistical Significance

- Everyone agrees that statistical significance of coefficients in a regression models is required
- However, statistical significance depends on sample size: anything can be significant provided we have large enough sample
- Thus, it may only be a necessary and not sufficient condition

Increase in c statistic

- Increase in the c statistic incurred with the addition of a new marker is not nearly as useful as the c statistic itself:
  - It has no intuitive interpretation
  - It is very small in magnitude when a few powerful risk factors are already in the model
  - It ignores the issue of calibration

Improving calibration?

- Quantifying change in calibration by decile chi-square will not work as they are not monotone to the number of risk factors
- None of the other measures seems to be

New metrics

- Some researchers argue that a performance metric quantifying usefulness of a new marker should be tied to the impact on clinical decision
- In some settings there exist meaningful cut-offs for assignment of risk categories
- For example, in cardiovascular field, risk >20% is considered high, <6% is considered low

Reclassification

- In these setting it might be useful to assess the degree of correct reclassification introduced by the new marker
- The Net Reclassification Improvement quantifies the amount or weighted percentage of correct reclassification
- Category dependent

Net Reclassification Improvement

- NRI* is calculated as a sum of two separate components: one for individuals with events and the other for individuals without events
- For events, we assign 1 for upward reclassification, -1 for downward and 0 for people who do not change their risk category
- The opposite is done for non-events
- We sum the individual scores and divide by numbers of people in each group

*Pencina, D'Agostino et al., Statist Med. 2008

Framingham Example: HDL

- 3264 women and men, 30-74 years of age, free of CVD followed for 10 years for the development of their first CHD event
- HDL cholesterol as the "new marker"
- Age, sex, diabetes, smoking, systolic BP, total cholesterol as the standard risk factors
- Cox PH models used for prediction; binary outcome used for assessment

NRI – Framingham HDL example

Events				
No HDL model	**Model with HDL**			
	< 6%	**6-20%**	**> 20%**	**Total**
< 6%	39 72.22	15 27.78	0 0.00	54
6-20%	4 3.81	87 82.86	14 13.33	105
> 20%	0 0.00	3 12.50	21 87.50	24
Total	43	105	35	183

Non-Events				
No HDL model	**Model with HDL**			
	< 6%	**6-20%**	**> 20%**	**Total**
< 6%	1959 93.24	142 6.76	0 0.00	2101
6-20%	148 16.78	703 79.71	31 3.51	882
> 20%	1 1.02	25 25.51	72 73.47	98
Total	2108	870	103	3081

NRI calculation

- NRI = (29 – 7)*(1/183)+(174-173)*(1/3081)
- NRI = 12.1%, p-value (asymptotic) < 0.0001
- Alternatively: Let p = 183/3264

  NRI = (1/p)*(29 – 7)/3264 +

  (1/(1-p))* (174-173)/3264
- Note:

  (1/p)/(1/(1-p)) = (1-p)/p = non-event odds

Integrated Discrimination Improvement

- Difference in discrimination slopes:

$$IDI = (\dot{p}_{+\ marker,\ events} - \dot{p}_{+\ marker,\ nonevents}) - (\dot{p}_{-\ marker,\ events} - \dot{p}_{-\ marker,\ nonevents})$$

$\dot{p}$ = mean predicted probability of event among events and nonevents, based on models with and without the new marker

HDL Example Results

- Hazard Ratio = 0.65, p-value < 0.0001
- AUC increase from 0.762 to 0.774, difference p-value = 0.092;
- NRI = 12.1%, p-value < 0.0001, almost entirely due to improvement in classification of events;
- Relative IDI = 7% (0.009 on the absolute scale with p-value = 0.008);

Diabetes example

	Sex only		Multivariable adjusted	
	-genetics score	+genetics score	-genetics score	+genetics score
Odds ratio 95% CI		1.12 1.07, 1.17		1.11 1.05, 1.17
C statistic P-value	0.534	0.581 0.01	0.900	0.901 0.49
NRI P-value		4.1% 0.004		2.1% 0.17

Summary

- Model building, development and validation is a process not a paper
- Model performance needs to be carefully examined and validated
- Different metrics may be needed for performance evaluation and assessment of new marker utility