Developing Risk Prediction Models Using Nested Case-Control Data

28 May 2015



- Recently, there has been an explosion of prediction models developed to predict risk of various diseases
- A simple "risk prediction models" search in PubMed reveals that there are 300 publications in the last 10 years alone.
- The number is rapidly growing, in a large extent due to discoveries of new biomarkers from 'omics fields.
- Among others, risk prediction models are useful for identifying and prioritizing high-risk individuals for interventions.

ATP III risk calculator

- The most widely-used risk prediction model is arguably the Adult Treatment Panel III (ATP-III) risk calculator for estimating 10-year risk of coronary heart disease
- It has been used as the basis of statin (LDL-lowering drug) for those with predicted risk > 20%.
- http://cvdrisk.nhlbi.nih.gov/calculator.asp

| | • 🛃 Search • 🔡 Share More 39 | | | | | |
|---|---|------------------------------|--|--|--|--|
| Tuesday, May 26, 2015 | | | | | | |
| Information for Health Professionals | Risk Assessment Tool for Estimating Y | our 10-year Risk of Having a | | | | |
| Clinical Practice Guidelines | Heart Attack | | | | | |
| Heart & Vascular Information Lung Information | The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and elder who do not have heart disease or diabetes. To find your risk score, enter your information the calculater below. | | | | | |
| Sleep Information | Age: | years | | | | |
| Interactive Tools and Resources | Gender: | ○ Female ○ Male | | | | |
| Education Campaigns | Total Cholesterol: | mg/dL | | | | |
| National Education Programs | HDL Cholesterol: | mg/dL | | | | |
| Continuing Education Opportunities | Smoker: | ○ No ○ Yes | | | | |
| Health Observances | Systolic Blood Pressure: | mm/Hg | | | | |
| | Are you currently on any medication to treat high blood pressure. | ○ No ○ Yes | | | | |
| | Calculate Your 10 | -Year Risk | | | | |
| | G110 | | | | | |

- The ATP III calculator is based on a risk prediction model developed using data from the Framingham cohort.
- Cohort study is used as alternatives such as case-control cannot unbiasedly estimate absolute risk.
- The underlying statistical model is the proportional hazard (PH) model.

(https://www.framinghamheartstudy.org/risk-functions/ coronary-heart-disease/hard-10-year-risk.php). • The hazard rate, $\lambda(t)$ = the probability of experiencing event at t, given that subject survives a moment before.

$$egin{array}{rcl} \lambda(t) &=& P(T\in[t,t+\delta t)|T\geq t-\delta t) \ &=& rac{f(t)}{S(t)} \end{array}$$

• Hazard rate depends on individuals' characteristics, i.e., individual with more risk factors should have larger hazard rates.

$$\lambda_i(t) = \lambda_0(t) \exp[x_i^T \beta + z_i^T \gamma]$$

- $\lambda_0(t)$ is called *baseline hazard function*
- β and γ are log hazard ratios and reflects the effect of exposure on hazard.

- Suppose we observe K unique event times at $t_1, t_2, \ldots t_K$ for individuals $1, 2, \ldots k$ respectively.
- The probability of observing an event occurred at t_i for individual i, given that event times are unique (there can only be one event at t_j) is,

$$\frac{\exp[x_i^T\beta + z_i^T\gamma]}{\sum_{j:j\in\Re_i}\exp[x_j^T\beta + z_j^T\gamma]}$$

• Where \Re_i is the riskset at time t_i (the subset of subjects that were still at risk moment before t_i).

 Assuming independence between the different events, the parameters are estimated by maximizing the partial likelihood,

$$L(\beta,\gamma) = \prod_{i=1}^{K} \frac{\exp[x_i^T \beta + z_i^T \gamma]}{\sum_{j:j \in \Re_i} \exp[x_j^T \beta + z_j^T \gamma]}$$

- Note:
- The log hazard ratios can be estimated without the need to specify the baseline hazard function.
- The likelihood is simply a product over all different event times.
- A particular individual can potentially be in multiple risk-set.

A quick review of PH model

- But, if we are interested in the absolute risk, we will need to estimate the baseline hazard function
- Given $\hat{\beta}$ and $\hat{\gamma}$, we use the Breslow estimator which assumes the cumulative baseline hazard function is a step function with 'jumps' at the unique event times,

$$\hat{\Lambda}_0(t) = \sum_{i; t_i \leq t} rac{1}{\sum_{j \in \Re_i} \exp[x_j^T \hat{eta} + z_j^T \hat{\gamma}]}$$

• The absolute risk is estimated as

$$\begin{aligned} \hat{F}_i(t) &= 1 - \hat{S}_i(t) \\ &= 1 - \exp[-\hat{\Lambda}_i(t)] \\ &= 1 - \exp[-\hat{\Lambda}_0(t)]^{\exp[x_i^T\hat{\beta} + z_i^T\hat{\gamma}]} \end{aligned}$$

- **Discrimination**: how well the model separate those who will develop the event from the rest.
- **Calibration**: how well the model estimates the *true* absolute risk. Calibration quality can be poor when models developed in one population is applied to another population with different event rates. H-L GoF test is often used to examine the calibration quality.

Discrimination Quality: AUC and C-statistic

 When there is no censoring (we know event times for everybody), Area Under the Curve (AUC) statistic measure the degree of concordant between the predicted risk and outcome (event time). Let (T_i, T_j) be event times for individuals i and j and F̂_i(t), F̂_j(t) the predicted risk,

$$AUC = P[T_i < T_j, \hat{F}_i(t) < F_j(t) \text{ OR } T_i > T_j, \hat{F}_i(t) > F_j(t)]$$

If c_{ij} , i < j is an indicator whether individuals (i, j) are concordant and there are *n* individuals in the cohort, then AUC is estimated as

$$A\hat{U}C = \frac{2}{n(n-1)}\sum_{i}\sum_{j>i}c_{ij}$$

Discrimination Quality: AUC and C-statistic

- With censoring, we cannot compare some of the pairs. These pairs are 'unuseable'
 - If both *i* and *j* are censored, then we do not know the ordering of event times.
 - Also, if we observe event for i at T_i but j already censored at this point.
- Harrell (1996), also Pencina and D'Agostino (2004) proposed to use only the 'useable' pairs and estimate what they call c-statistic,

$$\hat{C} = rac{1}{Q} \sum c_{ij}$$

where the summation is across all usable pairs and Q the number of such pairs.

 When comparing two models ('OLD' and 'NEW'), the 'NEW' is better if it has statistically higher c-statistic.

Discrimination Quality: AUC and C-statistic

- To improve the c-statistic, the new model needs to be able turn the discordant pairs into concordant ones.
- It has been observed that even a fairly large effect size of the new biomarkers will only result in a meagre increase in c-statistic (Pepe et al., 2004; Ware 2006).
- In clinical application, predicted risk are often categorized and recommendation will be based on categorized risk
- E.g, ATP III guideline categorized 10-year risk into <10%, 10 20% and >20%. Statin initiation is recommended only for those in the highest category and optional for those in the second.
- From this perspective, even a small change in the predicted risk will still be deemed useful if it re-classifies the individual into a more appropriate risk category.

Discrimination Quality: Net Reclassification Index (NRI)

- For those who developed event, appropriate re-classification occur if an individual is moved up in the risk category, by the new model
- For those who did not develop event, appropriate re-classification occur if an individual is moved down in the risk category, by the new model

Let

 $\hat{p}_{up}^{e} =$ prop of those with events whose risk classification is moved up $\hat{p}_{down}^{e} =$ prop of those with events whose risk classification is moved down $\hat{p}_{up}^{ne} =$ prop of those without events whose risk classification is moved up $\hat{p}_{down}^{ne} =$ prop of those without events whose risk classification is moved down • The estimated net reclassification for those with events is

$$\hat{NRI}^e = \hat{p}^e_{up} - \hat{p}^e_{down}$$

• The estimated net reclassification for those without events is

$$\hat{NRI}^{ne} = \hat{p}_{down}^{ne} - \hat{p}_{up}^{ne}$$

Pencina et al (2008) gives expression for the asymptotic variance of these estimates and hypothesis testing can thus be carried out.

Alternatives to Cohort Study

- Cohort study is often expensive to perform as measurements need to be collected on all cohort members.
- With limited research budget, we need a viable alternative.
- Two study designs that utilize only a sub-cohort: case-cohort (CCH) and nested case-control (NCC) studies.
- Difference: CCH selects the subset at baseline, NCC selects the "controls" post-baseline as event occurs (Langholz and Thomas, 1990 for details).
- CCH has been shown to unbiasedly estimate absolute risk [Ganna et al (2012) and Cook et al (2012)] but NCC has not...

• In fact, Ganna et al showed that matched NCC biasedly estimate absolute risk (Figure 1D from Ganna et al).



Agus Salim ()

VicBiostat Seminar, 28 May 2015

- To understand why they observe that bias, we look at the NCC sampling and how the parameters are estimated.
- The sampling of controls in NCC is based on incidence density sampling: for each incident case, we select a subset of the riskset.



• Note that probability of selection for each control depends on various factors (length of follow-up, matching factors etc).

Agus Salim ()

• Parameters are estimated by maximizing partial likelihood just like in the cohort study case, but the denominator is different,

$$L(\beta,\gamma) = \prod_{i=1}^{K} \frac{\exp[x_i^T \beta + z_i^T \gamma]}{\sum_{j:j \in R_i} \exp[x_j^T \beta + z_j^T \gamma]}$$

where R_i is the set that consist of the case (individual *i*) and the selected controls.

Problem 1: Epidemiologist like to match case and controls on confounders (for efficiency, logistic reasons). Suppose that the case and controls are matched so that the matched controls have the same z value as the case, z_j = z_i, ∀j ∈ R_i, then we lose the ability to estimate log HR associated with z,

$$L^{matched}(\beta) = \prod_{i=1}^{K} \frac{\exp[x_i^{T}\beta]}{\sum_{j:j\in R_i} \exp[x_j^{T}\beta]}$$

- Problem 2: even in the unmatched case, the Breslow estimator for cumulative baseline hazard will be biased, because we have higher proportion of cases in the NCC data relative to the proportion in the cohort.
- Langholz and Borgan (1997) attempted to remedy this by using a weighted version of Breslow estimator, where the controls selected at time t_i are weighted by the inverse of their probability of being selected

$$\hat{\Lambda}_0^{LB}(t) = \sum_{i; t_i \leq t} \frac{1}{\sum_{j \in \mathcal{R}_i} \frac{n(t_i)}{m+1} \exp[x_j^T \hat{\beta} + z_j^T \hat{\gamma}]}$$

• This is what was used by Ganna et al in the case of unmatched NCC study.

- But we have not solved the first problem.
- Samuelsen (1997) proposed an alternative approach to estimate hazard ratios from NCC data. His approach 'breaks' the case-control pairing and all unique individuals are placed in one pool...
- In the partial likelihood, each individual is weighted by the inverse of their probability of being included into the study (IPW). For cases, we can assume this probability is 1. The parameters are then estimated using a 'weighted' partial likelihood

$$L^{w}(\beta,\gamma) = \prod_{i=1}^{K} \frac{w_{i} \times \exp[x_{i}^{T}\beta + z_{i}^{T}\gamma]}{\sum_{j:j \in R_{i}} w_{j} \times \exp[x_{j}^{T}\beta + z_{j}^{T}\gamma]}$$

 R_i is the set of all individuals in the pool who has not experienced event moment before time t_i and w_j is the weight for individual j.

 For controls, the probability of inclusion can computed using K-M type formula

$$m{p}_j = 1 - \prod_{j:t_i < t_j} \left[1 - rac{m}{n(t_i) - 1}
ight]$$

- Note: the product is taken over all event times when individual *j* has not experienced event. In the case of matched NCC, the product is only taken over all event times when individual *j* has not experienced event and the case has the same matching characteristics.
- The estimates from this approach have been shown to be unbiased, provided that the probability of inclusion depends only on variables fully-observed in the cohort.

• The cumulative baseline hazard is estimated as

$$\hat{\Lambda}_{0}^{w}(t) = \sum_{i;t_{i} \leq t} \frac{1}{\sum_{j \in R_{i}} w_{j} \times \exp[x_{j}^{T} \hat{\beta} + z_{j}^{T} \hat{\gamma}]}$$

Weighted Net Reclassification Index (wNRI)

• We can also use the IPW weights to calculate the weighted NRI. For those with events the estimate is

$$w\hat{N}RI^{e}=\hat{p}_{up}^{w,e}-\hat{p}_{down}^{w,e}$$

For those without events the estimate is

$$w\hat{N}RI^{ne} = \hat{p}^{w,ne}_{down} - \hat{p}^{w,ne}_{up}$$

where p^w are the weighted proportions, with individual weight given by the IPW weight. We also derive the variance expression of the wNRI needed for performing hypothesis testing.

- We generated 500 independent simulated cohorts (of 50,000 subjects each) that closely mimic the characteristics of subjects in the Singapore Chinese Health Study (SCHS) cohort. All simulated variables have the same mean and covariance structure as the corresponding variables in the SCHS cohort.
- The event times are generated assuming proportional hazard model with constant baseline hazard.
- Random censoring time is generated as exponential r.v with rate = 0.05
- Results in \sim 96% of cohort members being censored and an average follow-up time \sim 20 years.
- For simplicity, we assume everybody enters the cohort at beginning of the study (no staggered entry).

- Within each simulated cohort, two types of NCC study is conducted:
 (1) without matching, (2) with gender-matching
- Log hazard ratios, cumulative baseline hazard and absolute risk are estimated using: (1) full cohort data, (2) NCC data via Samuelsen (weighted) method, (3) NCC data via L-B method.
- Average estimates and empirical standard errors are calculated over 500 realisations

Simulation Results: Log hazard ratios

| | L-B Approach | | Weighte | ed Approach | Cohort | | True | |
|---------|--------------|-------|---------|-------------|--------|-------|-------|--|
| | Est | SE | Est | SE | Est | SE | Value | |
| Age | 1.056 | 0.006 | 1.056 | 0.006 | 1.056 | 0.004 | 1.056 | |
| Gender | 1.616 | 0.136 | 1.617 | 0.133 | 1.605 | 0.102 | 1.599 | |
| Chol | 1.006 | 0.001 | 1.006 | 0.001 | 1.006 | 0.001 | 1.006 | |
| HDL | 0.977 | 0.003 | 0.977 | 0.003 | 0.977 | 0.002 | 0.977 | |
| SBP | 1.013 | 0.002 | 1.013 | 0.002 | 1.013 | 0.001 | 1.013 | |
| Antihyp | 1.337 | 0.122 | 1.342 | 0.120 | 1.336 | 0.088 | 1.331 | |
| Smoke | 1.720 | 0.153 | 1.723 | 0.157 | 1.719 | 0.110 | 1.710 | |

Table 1: Hazard ratio estimates and their standard errors for NCC studies without matching. Estimates are averages across 500 simulated cohorts.

Table 2: Hazard ratio estimates and their standard errors for NCC studies with gender-matching. Estimates are averages across 500 simulated cohorts.

| | L-B Approach | | Weighted Approach | | Cohort | | True |
|---------|----------------|-------|-------------------|-------|--------|-------|-------|
| | Est | SE | Est | SE | Est | SE | Value |
| Age | 1.056 | 0.005 | 1.056 | 0.005 | 1.056 | 0.004 | 1.056 |
| Gender | Not Estimeable | NA | 1.600 | 0.119 | 1.604 | 0.104 | 1.599 |
| Chol | 1.006 | 0.001 | 1.006 | 0.001 | 1.006 | 0.001 | 1.006 |
| HDL | 0.977 | 0.003 | 0.977 | 0.003 | 0.977 | 0.002 | 0.977 |
| SBP | 1.013 | 0.002 | 1.013 | 0.002 | 1.013 | 0.001 | 1.013 |
| Antihyp | 1.342 | 0.114 | 1.339 | 0.112 | 1.333 | 0.083 | 1.331 |
| Smoke | 1.727 | 0.153 | 1.730 | 0.148 | 1.720 | 0.114 | 1.710 |

Simulation Results: Log hazard ratios

- Without matching, both L-B and weighted approaches unbiasedly estimate log hazard ratios
- With gender-matching, L-B approach unable to estimate log HR for gender
- There is noticeable benefit of matching in reducing the SE of estimates (better efficiency)

Simulation Results: Absolute Risk Estimates



Agus Salim ()

Simulation Results: Absolute Risk Estimates



Agus Salim ()

Application: Does CRP, creatinine and HbA1c improves CHD risk prediction?

- In western populations high-sensitivity C-reactive protein (hsCRP), and to a lesser degree serum creatinine and haemoglobin A1c (HbA1c) predict risk of coronary heart disease (CHD).
- Data on Asian populations where CHD are expected to increase are sparse and it is not clear if these biomarkers will actually improve the CHD risk classification.
- Study design: NCC with up to 3 controls per case, conducted within SCHS cohort. Controls are matched to case on age (+/- 1 years), gender, dialect group, blood storage time (+/- 6 months).
- Outcome: CHD (ICD 9:410-414). To be eligible for selections into NCC study, participants need to not have coronary heart disease or stroke, at the time of their blood collection.
- Study period: 1993-2010, average follow-up time: 12.4 years.

Application: Does CRP, creatinine and HbA1c improves CHD risk prediction?

- We will compare the model with only ATP III risk factors: age, total cholesterol, HDL-C, SBP, antihypertensive treatment and current smoking status with a new model with CRP, creatinine and HbA1c added.
- Parameters estimation used the weighted approach and separate analyses are conducted for male and female.
- I will discuss only results for male (298 cases, 667 controls).

Application: Basic Characteristics

Table 1: Characteristics of Study Subjects[†]

| | Men (N | = 965) | Women (N = 528) | | |
|-----------------------------------|---------------|----------------|-----------------|----------------|--|
| | Cases (N=298) | Control(N=667) | Cases (N=143) | Control(N=385) | |
| Age at blood collection (years) | 65.35(7.83) | 62.98(7.53) | 69.23(7.52) | 62.96(7.00) | |
| Total cholesterol (mg/dL) | 202.39(36.49) | 195.86(33.58) | 215.94(39.98) | 210.86(33.28) | |
| HDL-C (mg/dL) | 47.37(9.79) | 49.41(11.43) | 55.55(12.72) | 58.78(12.46) | |
| Systolic blood pressure (mm Hg) | 145.62(22.66) | 136.28(19.87) | 148.84(23.34) | 133.75(21.08) | |
| Anti-hypertensive treatment | 0.37(0.48) | 0.26(0.44) | 0.33(0.47) | 0.27(0.44) | |
| Currentsmoking | 0.42(0.49) | 0.31(0.46) | 0.16(0.37) | 0.05(0.21) | |
| hsCRP(mg/L) | 2.15(2.18) | 1.34(1.48) | 1.97(2.06) | 1.53(1.67) | |
| Serum creatinine (mg/dL) | 0.95(0.43) | 0.86(0.16) | 0.68(0.28) | 0.62(0.14) | |
| HbA1c(%) | 5.75(0.42) | 5.69(0.44) | 5.79(0.35) | 5.70(0.34) | |
| 10-year CHD event rate (per 1000) | 42 | | 25 | | |

[†] Figures are weighted means (weighted SD), calculated using nested case-control data with inverse of probability of inclusions as sampling weights.

Application: Model Comparisons

Table 2: Performance of Different Predictive Models for Men

Agus Salim ()

| | Hazard Ratio | | C-statistic | Cases NRI | Controls NRI |
|---------------------------------------|---------------------|---------|--------------|--------------|---------------|
| Model | (95% CI) | P-value | (SE) | (p-value) | (p-value) |
| M0: ATPIII | - | | 0.679(0.018) | | |
| Models with single biomarkers | | | | | |
| M1:ATPIII+log(CRP) | 1.254(1.055; 1.491) | 0.010 | 0.689(0.018) | 0.032(0.020) | 0.002(0.759) |
| M2:ATPIII + log(creatinine) | 4.817(2.101;11.044) | <0.001 | 0.685(0.018) | 0.047(0.012) | -0.011(0.136) |
| M3:ATPIII+log(HbA1c) | 1.825(0.207;16.057) | 0.588 | 0.679(0.018) | 0.000(1.000) | -0.002(0.483) |
| Models with multiple biomarkers | | | | | |
| M4: ATPIII+ log (CRP)+log(creatinine) | | | | | |
| log(CRP) | 1.225(1.030; 1.456) | 0.021 | 0.696(0.018) | 0.063(0.001) | -0.009(0.275) |
| log(creatinine) | 4.284(1.889;9.715) | <0.001 | 0.696(0.018) | 0.063(0.001) | -0.009(0.275) |

The weighted NRI were performed using ATP III risk cut-off (<10%, 10 - 20% and >20%).

VicBiostat Seminar, 28 May 2015

Application: SCHS Male



Application: Absolute risk estimates

Addition of CRP and creatinine increases predicted risk for high-risk individuals with elevated CRP and creatinine, while reducing the predicted risk for those with low CRP and creatinine



- Developing risk prediction model using NCC data is feasible and offers significant savings.
- Including CRP and serum creatinine into ATP III model improves CHD risk classification in Chinese and could potentially lead to better treatment recommendation for subjects with elevated CRP and creatinine not previously 'picked' by ATP III model.

Collaborators:

- National University of Singapore: Rob van Dam, Tai Eshyong
- **Singapore Chinese Health Study:** Koh Woon Puay (Duke-NUS), Yuan Jian-Min (U of Pittsburgh)
- The Australian National University: Alan Welsh

Funding:

- Singapore Chinese Health Study: National Institute of Health, USA [Grant Number NCI RO1 CA055069, R35 CA053890, R01 CA080205, R01 CA098497 and R01 CA144034]
- This work: Singapore National Medical Research Council [Grant Number 1261 (to Rob van Dam), 1270 (to Agus Salim)].

References

- Cai T, Zheng Y.(2012). Evaluating Prognostic Accuracy of biomarkers under Nested Case-control studies. Biostatistics, 13:89-100.
- Cook NR et al. (2012). Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. Circulation, 125:1748-1756.
- Ganna A et al. (2012). Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. American Journal of Epidemiology, 175:715-24.
- Harrell FE et al. (1996). Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15:361 387.
- Langholz B, Borgan O. (1995). Estimation of absolute risk from nested case-control data. Biometrics, 53: 767-774.
- Langholz B, Thomas DC. (1990). Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. American Journal of Epidemiology, 131:169-76.

- Pencina MJ, D'Agostino RB. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine, 23:21092123.
- Pencina MJ et al. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in Medicine, 27: 157-172.
- Pepe MS et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology, 159(9):882-90.
- Samuelsen, SO. (1997). A pseudolikelihood approach to analysis of nested case-control studies. Biometrika, 84: 379-394.
- Salim et al. Absolute risk estimation using nested case-control data. submitted.
- Salim et al. C reactive protein and serum creatinine, but not HbA1c, improve prediction of coronary heart disease risk in non-diabetic Chinese: the Singapore Chinese Health Study. submitted.
- Zhou QM, Zheng Y, Cai T. Assessment of biomarkers for risk prediction with nested case-control studies. (2013). Clinical Trials, 10:677-679.