SUPPLEMENTARY MATERIAL

associated with

"Quantifying the Value of Biomarkers for Predicting Mortality"

Dana A. Glei and Noreen Goldman

MEASURES OF DISCRIMINATION

Area under the receiver operating characteristic curve (AUC)

For the purposes of evaluating the incremental value of a marker, the AUC has some drawbacks. In particular, ΔAUC is insensitive to the inclusion of a novel biomarker if the baseline model possesses good discrimination, even if the effect size is large. Furthermore, ΔAUC does not take into account the magnitude of the difference in probabilities between models [1]; it considers the rank order of cases and noncases rather than the actual predicted probabilities [2]. In an effort to address these and other criticisms of the AUC, researchers have developed alternative measures of discrimination based on reclassification methods.

Net Reclassification Improvement (NRI)

The Net Reclassification Improvement (NRI) uses reclassification tables constructed separately for individuals that experience the event and those that do not [3]. It then quantifies the correct movement between risk categories (i.e., upwards for those with the event and downwards for who do not have the event). One drawback of the NRI is that it requires meaningful risk categories *a priori*, and the results are sensitive to the choice of categories [3, 4]. A newer category-free version, NRI(>0), addresses this issue by redefining upward and downward movement based on changes in the predicted probabilities: upward for those who died and downward for those who survived [5]. One can think of the NRI(>0) as a limiting case of the category-based NRI where each unique predicted probability represents its own category [6]. The NRI(>0) represents a summary measure of the correct upward versus downward movement in model-based probabilities for events and non-events [6].

The NRI(>0) is calculated as:

$$\boldsymbol{NRI}(>\mathbf{0}) = \mathbf{2} * \left\{\boldsymbol{P}\left(\boldsymbol{q}_{\text{New, Event}} > \boldsymbol{q}_{\text{Old, Event}}\right) - \boldsymbol{P}\left(\boldsymbol{q}_{\text{New, Non-Event}} > \boldsymbol{q}_{\text{Old, Non-Event}}\right)\right\} \\ = 2 * \left\{P(\text{Up}|\text{Event}) - P(\text{Up}|\text{Non-Event})\right\}, \tag{1}$$

where $q_{\text{New, Event}}$ and $q_{\text{Old, Event}}$ represent the predicted probability of the event among those who did in fact experience the event based on the "new" and "old" models, respectively; $q_{\text{New, Non-Event}}$ and $q_{\text{Old, Non-Event}}$ denote the corresponding probabilities among those who did not experience the event. $P(\text{Up}|\text{Event})$ represents the probability that $q_{\text{New, Event}}$ is greater than $q_{\text{Old, Event}}$, while $P(\text{Up}|\text{Non-Event})$ is the corresponding quantity among those who did not have the event. Thus, the NRI(>0) is the difference between the probability of upward movement for the two groups multiplied by two.

Among the discrimination measures discussed here, Pencina et al. [6] argue that the NRI(>0) is the best indicator of the true discriminatory potential of the added marker; unlike the AUC, the NRI depends mainly on the effect size of the added predictor rather than the strength of the baseline model. Thus, it addresses one of the major criticisms of the AUC, but it still does not take into account the magnitude of movements: it focuses only on net numbers with altered risk [6]. Consequently, as Cook [7] shows, one can get some anomalous results; for example, a new model may look worse because there are more changes in the wrong direction even though the incorrect changes are smaller than the correct changes. Pencina et al. [6] point out that in cases where there is some minimum change in risk that would be considered clinically meaningful, it may be preferable to calculate NRI(>*x*), where *x* represents that minimal change.

Integrated Discrimination Improvement (IDI)

Unlike the AUC and NRI, the IDI incorporates information about the magnitudes of changes in probabilities by weighting the movements by their magnitudes. The IDI is calculated as:

$$\text{IDI} = \underbrace{\left[\bar{\boldsymbol{q}}_{\text{New, Event}} - \bar{\boldsymbol{q}}_{\text{New, Non-Event}}\right]}_{\text{Slope (New Model)}} - \underbrace{\left[\bar{\boldsymbol{q}}_{\text{Old, Event}} - \bar{\boldsymbol{q}}_{\text{Old, Non-Event}}\right]}_{\text{Slope (Old Model)}}, \tag{2}$$

where $\bar{q}_{\text{New, Event}}$ and $\bar{q}_{\text{New, Non-Event}}$ represent the mean predicted probabilities of an event based on the "new" model for those who had the event and those who did not, respectively; $\bar{q}_{\text{Old, Event}}$ and $\bar{q}_{\text{Old, Non-Event}}$ denote the corresponding means based on the "old" model. The IDI can be directly interpreted as the amount of increase in the difference between the mean predicted probability of events and non-events [6].

Like ΔAUC, the IDI also represents a measure of overall improvement in sensitivity and specificity, but whereas the AUC weights cutoffs associated with high sensitivity more heavily, the IDI assigns equal weight to all values of sensitivity [3]. The mean probability of an event among those who experienced the event ($\bar{q}_{\text{Event}}$) represents the average sensitivity, whereas the mean probability of an event among those who did not experience the event ($\bar{q}_{\text{Non-Event}}$) can be viewed as the average of 1-specificity. Thus, if we rewrite Equation (2) as:

$$\text{IDI} = \underbrace{\left[\bar{\mathbf{q}}_{\text{New, Event}} - \bar{\mathbf{q}}_{\text{Old, }\textbf{Event}}\right]}_{\text{Change in Sensitivity}} - \underbrace{\left[\bar{\mathbf{q}}_{\text{New, Non-Event}} - \bar{\mathbf{q}}_{\text{Old, Non-Event}}\right]}_{\text{Change in (1-Specificity)}}, \tag{3}$$

we see that the IDI can be interpreted as the difference between improvement in average sensitivity and any potential increase in the average of 1-specificity.

The IDI bridges the perspectives of the ΔAUC, which depends heavily on the strength of the baseline model, and NRI(>0), which is the least dependent on the baseline model strength [6]. Whether such dependence is desirable or not is debatable. Kerr et al. [9] argue that invariance to the strength of the baseline model is not necessarily desirable: if the baseline model is almost perfect, then the incremental value of any additional marker should be small. Pencina et al. [6, 10] contend that the preferred metric depends on one's purpose: the AUC is preferred when the focus is on the model itself rather than the variables to be added, whereas the NRI(>0) is better for assessing the true discriminatory potential of a new marker compared with other markers (i.e., marginal strength). The IDI falls somewhere in between.

The IDI differs from the AUC and the NRI(>0) in two additional ways. First, the IDI takes into account the magnitude of changes in the probabilities, whereas the AUC and NRI(>0) are based only on the net numbers with altered risk. Second, the IDI depends on the event rate in a way that the other measures do not. Thus, it is more heavily influenced by model calibration (i.e., the ability to correctly estimate the probability of an event) and cannot be compared across studies with different event rates.

PREDICTED PROBABILITY OF DYING BY THE END OF FOLLOW-UP

To calculate the AUC using ROC analysis, we use the model coefficients to compute the predicted probability of dying by June 30, 2011 for each respondent, which is then compared with the observed outcome (i.e., whether or not the respondent actually died). The Gompertz proportional hazards model takes the following form:

$$\log \boldsymbol{\lambda}(\boldsymbol{t}) = \boldsymbol{x\beta} + \boldsymbol{\gamma t}\,, \tag{4}$$

where t represents time measured in age, $\lambda(t)$ is the hazard rate at time t (age), γ denotes the age slope, x represents a covariate, and β is the corresponding regression coefficient. In our case, we fit a model that allows for non-proportional hazards. That is, γ is a function of x (i.e., the covariate x is interacted with age)

$$\boldsymbol{\gamma} = \boldsymbol{\gamma_0} + \boldsymbol{\gamma_1 x}\,. \tag{5}$$

For this model, the conditional probability of surviving from the date of the survey (t_0) to the end of follow-up (t_1) can be computed as:

$$\boldsymbol{S}(\boldsymbol{t_1}|\boldsymbol{t_0}) = \boldsymbol{exp}\left\{-\boldsymbol{e^{x\beta}}(\boldsymbol{e^{\gamma t_1}} - \boldsymbol{e^{\gamma t_0}})/\boldsymbol{\gamma}\right\}. \tag{6}$$

Thus, for each respondent, we: a) calculate the linear prediction ($x\beta$) based on the observed value(s) for the covariate(s) and the model coefficient(s); b) compute γ given the observed value(s) of any covariates that were interacted with time t (age); and c) estimate conditional survival for using Equation (6).

The probability of dying between t_0 and t_1 is simply the complement:

$$\widehat{\boldsymbol{q}}(\boldsymbol{t_1}|\boldsymbol{t_0}) = \boldsymbol{1} - \widehat{\boldsymbol{S}}(\boldsymbol{t_1}|\boldsymbol{t_0})\,. \tag{7}$$

Table S1. Descriptive Statistics for Social and Demographic Characteristics, Self-reported Indicators of Health Status, and Survival Status, Unweighted Analyses, Taiwan, 2006-2011, SEBAS

	Analysis sample (N=639)
Social and demographic characteristics	
Age at the 2006 exam (60-97), mean (SD)	73.1 (7.4)
Female, %	43.6
Mainlander, %	16.1
Urban resident, %	54.3
Years of completed education (0-17), mean (SD)	5.5 (4.6)
Social integration (-1.5 to 1.6), mean (SD)[a]	0.1 (0.5)
Perceived availability of social support (0.5-4.0), mean (SD)[b]	3.1 (0.7)
Self-reported health indicators (2006)	
Self-assessed health status (1-5, 5=excellent), mean (SD)	3.0 (1.0)
Index of mobility limitations (-0.7 to 3.2), mean (SD)[c]	0.8 (1.3)
History of diabetes, %	18.8
History of cancer, %	5.3
Number of hospitalizations in the past 12 months (0-11), mean (SD)	0.3 (0.8)
Smoking status	
Never, %	57.8
Former, %	23.8
Current, %	18.5
Died between the 2006 exam and December 31, 2011, %	16.3

[a] This index was created by standardizing each of 10 indicators (network size, network range, married/partnered, household size, does not live alone, number of friends, religious attendance, socializing with others, volunteer work, participation in social organizations) from the 2003 Taiwan Longitudinal Study of Aging (TLSA) and then calculating the mean across valid items if at least eight items were valid (α=0.72). See Table S3 for details.

[b] Each of the following indicators was coded 0-4: family/friends willing to listen; family/friends make you feel cared for; satisfaction with emotional support received from family; can count on family to take care of you when you are ill. We calculated the mean across valid items if at least 3 items were valid (α=0.84).

[c] Each of eight tasks was coded on a four-point scale (0=no difficulty, 1=some difficulty, 2=great difficulty, 3=unable): stand for 15 minutes, squat, raise both hands overhead, grasp or turn objects with his or her fingers, lift or carry an object weighing 11-12kg, walk 200-300m, run 20-30m, and climb two or three flights of stairs. Based on the recommendations of Long and Pavalko [11], we summed the eight items (potential range 0-24), added a constant (0.5), and took the logarithm of the result to denote relative effects.

Table S2. Summary Statistics for Individual Biomarkers and Changes in Biomarkers, Unweighted Analyses, Taiwan, 2000-2006, SEBAS (N=639)

			Mean (SD) for the Transformed Markers:	
	Units	Transformation	Level in 2006	Change (2006 – 2000)
Systolic blood pressure (SBP)	mmHg	log	4.90 (0.15)	-0.01 (0.16)
Diastolic blood pressure (DBP)	mmHg	log	4.27 (0.15)	-0.13 (0.15)
High-density lipoprotein cholesterol (HDL)	mg/dL	log	3.83 (0.27)	-0.02 (0.22)
Ratio of total to HDL cholesterol (TC/HDL)	ratio	log	1.44 (0.27)	0.00 (0.23)
Triglycerides	mg/dL	log	4.57 (0.51)	-0.08 (0.42)
Glycosylated hemoglobin (HbA1c)	%	$-1/(\mathrm{HbA1c})^2$	-0.03 (0.01)	0.01 (0.01)
Body Mass Index (BMI)	$\frac{\mathrm{weight(kg)}}{(\mathrm{height(m)})^2}$	log	3.19 (0.15)	0.00 (0.08)
Waist circumference	cm	none	84.89 (9.93)	-0.55 (5.97)
Interleukin-6 (IL-6)	pg/mL	log	1.06 (0.79)	0.26 (0.89)
C-reactive protein (CRP)	mg/L	log	-2.01 (1.12)	0.52 (1.50)
Soluble intercellular adhesion molecule 1 (sICAM-1)	ng/mL	square root	16.53 (2.87)	1.11 (2.28)
Soluble E-selectin (sE-selectin)	ng/mL	log	3.57 (0.58)	-0.17 (0.43)
Dehydroepiandrosterone sulfate (DHEAS)	µg/dL	square root	8.87 (3.21)	0.25 (2.07)
Cortisol	µg/g	log	2.68 (0.87)	-0.28 (0.96)
Epinephrine	µg/g	log	1.25 (0.58)	0.12 (0.62)
Norepinephrine	µg/g	log	3.17 (0.53)	0.21 (0.53)
Creatinine Clearance (CrCl)	ml/min	none	58.23 (19.88)	-5.13 (11.11)
Albumin	g/dL	cubed	83.59 (17.25)	-9.04 (15.19)
Homocysteine (Hcy)	µmol/L	log	2.48 (0.39)	-0.20 (0.31)

Table S3. Index of social integration: description and coding of each component

Indicator	Definition	Coding
Network size	Number of friends and relatives with whom the respondent lives or has regular contact	Recoded <5, 5-7, 8-10, 11-14, 15-19, 20-29, 30+.
Network range	Number of types of relationships in social network	One point each for spouse/partner, kids, other relatives, non-relatives; range=0-4.
Married/partner	Dummy indicating that the respondent is married or lives with a companion.	
Household size		
Does not live alone	Dummy indicating that the respondent does not live alone.	
Number of friends	Number of close friends and neighbors with whom the respondent has weekly contact	Recoded 0, 1-2, 3-4, 5-9, 10-19, 20+.
Religious attendance	How often the respondent attends church or temple	Response categories: never, rarely, sometimes, often.
Socializing	How often the respondent socializes with friends, neighbors, or relatives.	Response categories: never, less than once a month, two to three times a month, once or twice a week, nearly daily.
Volunteer work	Dummy indicating that the respondent does volunteer work.	
Participation in social organizations	Whether respondent participates in the following activities/organizations: 1) Group activities (e.g., singing, dancing, tai chi, or karaoke) 2) Neighborhood association (e.g., women's association or arts & crafts classes) 3) Religious organization (e.g., church or temple committee) 4) Occupational associations for farmers, fishermen, or other professional group, civic group, Lion's Club, etc. 5) Political association (e.g., political party) 6) Social service groups (e.g., Lifeline, relief association, benevolent societies, charities, etc.) 7) Village or lineage association 8) Elderly club (e.g., Elderly Association, Evergreen Recreation Club, etc.)	One point for each type of organization in which the respondent participates; range = 0-7.

Table S4. Log Likelihood (*L*), *L* ratio test, and Akaike's Information Criterion (AIC) for Models Predicting Mortality as a Function of Social and Demographic Characteristics, Self-reported Indicators of Health Status, and Biomarkers, Taiwan, 2006-2011, SEBAS (N=639)

			L ratio test [c]			
Model	Description	log *L*	χ^2	df	*p* value	AIC [c]
1	Baseline: Self-reported indicators only[a]	-83.6				
			vs. Model 1			
2	Model 1 + 19 Individual biomarkers (2006)[b]	-52.5	62.2	19	<0.001	177.0
4a	Model 1 + 8 Cardiovascular/metabolic markers (2006 and changes 2000-06)	-71.4	24.5	16	0.080	208.8
4b	Model 1 + 4 Inflammatory markers (2006 and changes 2000-06)	-65.1	37.0	8	<0.001	180.3
4c	Model 1 + 4 Neuroendocrine markers (2006 and changes 2000-06)	-74.1	19.0	9	0.025	200.3
5a	Model 1 + SBP (2006 and change 2000-06)	-79.8	7.8	2	0.021	197.5
5b	Model 1 + DBP (2006 and change 2000-06)	-82.5	2.2	2	0.327	203.1
5c	Model 1 + HDL (2006 and change 2000-06)	-80.7	5.9	2	0.052	199.4
5d	Model 1 + TC/HDL (2006 and change 2000-06)	-82.1	3.0	2	0.220	202.3
5e	Model 1 +Triglycerides (2006 and change 2000-06)	-81.8	3.8	2	0.153	201.5
5f	Model 1 + HbA1c (2006 and change 2000-06)	-82.7	1.9	2	0.386	203.4
5g	Model 1 + BMI (2006 and change 2000-06)	-81.4	4.6	2	0.102	200.7
5h	Model 1 + Waist circumference (2006 and change 2000-06)	-82.1	3.1	2	0.207	202.1
5i	Model 1 + IL-6 (2006 and change 2000-06)	-70.9	25.4	2	<0.001	179.9
5j	Model 1 + CRP (2006 and change 2000-06)	-79.8	7.7	2	0.021	197.6
5k	Model 1 + sICAM-1 (2006 and change 2000-06)	-75.2	16.8	2	<0.001	188.5
5l	Model 1 + sE-selectin (2006 and change 2000-06)	-78.6	10.1	2	0.006	195.2
5m	Model 1 + DHEAS (2006 and change 2000-06)	-78.1	11.0	3	0.012	196.3
5n	Model 1 + Cortisol (2006 and change 2000-06)	-82.7	1.9	2	0.379	203.4
5o	Model 1 + Epinephrine (2006 and change 2000-06)	-81.6	4.1	2	0.131	201.2
5p	Model 1 + Norepinephrine (2006 and change 2000-06)	-82.8	1.6	2	0.445	203.7
5q	Model 1 + CrCl (2006 and change 2000-06)	-80.5	6.2	2	0.045	199.1
5r	Model 1 + Albumin (2006 and change 2000-06)	-80.6	6.0	2	0.049	199.3
5s	Model 1 + Hcy (2006 and change 2000-06)	-73.3	20.7	2	<0.001	184.6
			vs. Model 2			
3	Model 2 + Changes in 19 Individual biomarkers (2006 and changes 2000-06)[b]	-36.4	32.3	20	0.041	184.8

[a] Baseline model adjusts for: age (time-scale), sex, Mainlander, urban, education, social integration, perceived availability of support, smoking status, self-assessed health status, index of mobility limitations, history of diabetes, history of cancer, and number of hospitalizations in the past 12 months.

[b] The 19 biomarkers include cardiovascular/metabolic (SBP, DBP, ratio TC/HDL, HDL, triglycerides, HbA1c, BMI, waist), inflammatory (IL-6, CRP, sICAM-1, sE-selectin), and neuroendocrine markers (DHEAS, cortisol, epinephrine, norepinephrine) along with a few other markers that are unrelated biologically (creatinine clearance, serum albumin, homocysteine).

[c] Values for the *L* and AIC are based on the average across five multiply imputed datasets.

REFERENCES

1. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008**; 54(1)**: 17-23. DOI: 10.1373/clinchem.2007.096529.

2. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007**; 115(7)**: 928-935. DOI: 10.1161/CIRCULATIONAHA.106.672402.

3. Pencina MJ, D'Agostino RB,Sr., D'Agostino RB,Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008**; 27(2)**: 157-72; discussion 207-12. DOI: 10.1002/sim.2929.

4. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009**; 150(11)**: 795-802.

5. Pencina MJ, D'Agostino RB S, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011**; 30(1)**: 11-21. DOI: 10.1002/sim.4085.

6. Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACW, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol* 2012**; 176(6)**: 473-481. DOI: 10.1093/aje/kws207.

7. Cook NR. Clinically relevant measures of fit? A note of caution. *Am J Epidemiol* 2012**; 176(6)**: 488-491. DOI: 10.1093/aje/kws208.

8. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Am J Epidemiol* 2012**; 176(6)**: 482-487. DOI: 10.1093/aje/kws210.

9. Pencina MJ, D'Agostino RB, Demler OV, Janssens AC, Greenland P. Pencina et al. respond to "The incremental value of new markers" and "Clinically relevant measures? A note of caution". *Am J Epidemiol* 2012**; 176(6)**: 492-494. DOI: 10.1093/aje/kws206.

10. Long JS, Pavalko E. Comparing alternative measures of functional limitation. *Med Care* 2004**; 42(1)**: 19-27. DOI: 10.1097/01.mlr.0000102293.37107.c5.