Quantifying the Value of Biomarkers for Predicting Mortality

Noreen Goldman[a]

Dana A. Glei[b,*]

Maxine Weinstein[b]

[a] Office of Population Research, Princeton University
[b] Center for Population and Health, Georgetown University

*** Correspondence to:** Dana A. Glei, 5985 San Aleso Court, Santa Rosa, CA, 95409-3912, USA (e-mail: dag77@georgetown.edu). Phone: 1 (707) 539-5592. Fax: 1 (707) 978-3213.

ABSTRACT

Objective: We apply three discrimination measures to evaluate the incremental value of biomarkers – above and beyond self-reported measures – for predicting all-cause mortality and assess whether all three lead to the same conclusions.

Study Design and Setting: We use longitudinal data from a nationally representative sample of older Taiwanese (*n* = 639, aged 54+ in 2000, examined in 2000 and 2006, with mortality follow-up through 2011). We estimate age-specific mortality using a Gompertz hazard model.

Results: The broad conclusions are consistent across the three discrimination measures: (1) inclusion of 19 biomarkers substantially enhances survival prediction compared with self-reports alone; (2) incorporating changes (2000-06) in biomarkers yields a moderate improvement over one-time measurement; and (3) inflammatory markers offer stronger prediction than either cardiovascular/metabolic or neuroendocrine measures. Although the rank ordering of individual biomarkers varies across the three measures, the following is true for all three: interleukin-6 is the strongest predictor, the other three inflammatory markers make the top 10, and homocysteine ranks second or third.

Conclusion: The degree of consistency across metrics appears to vary with the level of detail inherent in the research question. Researchers would be wise to confirm findings with multiple discrimination measures before claiming victory.

INTRODUCTION

Epidemiologists and clinicians have a long-standing interest in identifying biomarkers that have prognostic value in predicting health events or death. Here we broaden this conventional inquiry by examining the incremental value of an extensive set of biomarkers – both standard and non-clinical markers – in predicting survival, above and beyond questionnaire-based information on self-reports of health and socio-demographic information. Since the 1990s, an increasing number of population-based social surveys have collected biological measures alongside detailed household questionnaires that obtain information on health and disability. Unfortunately, evaluations of the usefulness of these data collection efforts are seriously lacking, a critical concern in light of the financial costs, logistic complications, respondent burden, ethical concerns and threats to privacy of the data.

Determining the incremental value of biomarkers for risk assessment is not straightforward. In particular, statistical significance is insufficient because it is strongly influenced by sample size and because it fails to capture substantive importance. For such evaluations, epidemiologists and clinicians have relied primarily on measures of discrimination – that is, determining how well a regression model distinguishes individuals who experience an event from those who do not. The measure used most frequently is the area under the receiver-operating-characteristic curve (AUC). Because the focus of research has generally been on the *incremental* value of biomarkers, the corresponding metric has been the change in AUC (ΔAUC) attributable to the biomarker(s) of interest.

A serious limitation of ΔAUC – its strong dependence on the strength of the baseline model – has resulted in very small improvements in the AUC when a marker has been added to a baseline model that discriminates very well (e.g., AUC > 0.80). In response to this limitation, alternative methods have been proposed to compare predictive risk models. What remains unclear is the extent to which different measures yield similar conclusions. Pencina and colleagues underscore the distinct strengths and weaknesses of various discrimination measures and argue for presentation of multiple measures when assessing the incremental predictive value of novel markers [1]. Steyerberg and colleagues demonstrate that different discrimination measures favor the inclusion of different marker [2]. Such recommendations and findings are disturbing because there are no guidelines for identifying the preferred metric: these discrimination measures do not have any clinical interpretation or clinically-based cutoff values [3] and consensus for quantifying improvements in risk prediction appears to be lacking [4]. In addition, how the measures rank various models is likely to depend on the particular research question.

In this paper we apply three discrimination measures to evaluate the prognostic value of a set of biomarkers that were collected in the Social Environment and Biomarkers of Aging Study (SEBAS) in Taiwan [5], a pioneering survey of older adults. Similar arrays of markers have been included in other recent biosocial surveys. We consider four questions related to the value of adding biomarkers to models predicting five-year survival. First, do biomarkers have incremental value after adjusting for extensive self-reported information? Second, do *changes* in biomarker values provide better discrimination than a one-time measurement? Third, which cluster of biomarkers – standard cardiovascular/metabolic, inflammatory, or neuroendocrine – provides the strongest prediction? Finally, we pose a more nuanced question: which individual biomarkers are the strongest predictors?

An analysis of similar questions based only on the AUC was recently published [6]. Our objective is to assess the extent to which two additional discrimination measures provide consistent answers to these important issues.

METHODS

Data

Our data come from a cohort study in Taiwan, a population whose life expectancy and cause of death structure are similar to those observed in other industrialized countries including the US [7-9]. The SEBAS cohort was based on a nationally representative random sample of Taiwanese aged 54 and older in 2000. In 2000, in-home interviews were completed with 1,497 respondents, 1,023 of whom also completed a physical examination. Exam participants did not differ significantly from nonparticipants in ways likely to introduce serious bias [10]. Six years later, a follow-up was conducted with those who completed the 2000 exam and survived to 2006: Of the 846 who survived, 757 completed the in-home interview and 639 participated in the physical examination. The study protocol was approved by human subjects committees in Taiwan, Georgetown University and Princeton University. A public use dataset is available at http://www.icpsr.umich.edu/icpsrweb/NACDA/studies/3792/version/7.

The physical examination followed a similar protocol in both waves. Several weeks after the household interview, participants collected a 12-hour overnight urine sample (7pm to 7am), fasted overnight, and visited a nearby hospital the following morning for a physical examination. Union Clinical Laboratories in Taipei analyzed the blood and urine specimens; a sample of duplicate specimens was sent to Quest Diagnostics in the US for comparison. Details regarding response rates, sample attrition, exam participation, intra-lab reliability, inter-lab correlations, and compliance with the medical protocol are provided elsewhere [11].

Survival status as of January 1, 2012 was ascertained by linkage to the death certificate file maintained by the Taiwan Department of Health and to the household registration database maintained by the Ministry of the Interior. The analysis sample was based on the longitudinal cohort that completed the exam in both 2000 and 2006 (n=639, 104 of whom died by December 31, 2011). The mean length of mortality follow-up was 5.1 years. Because 89 of the respondents had missing data for at least one covariate, we followed standard practices of multiple imputation [12, 13] based on five imputed datasets to handle missing data.

Biomarkers and Control Variables

Biomarkers. We include 19 biomarkers that have been shown by prior studies to be associated with all-cause mortality. They comprise three clusters of biologically-related markers: 1) eight standard cardiovascular/metabolic risk factors—systolic and diastolic blood pressure, high-density lipoprotein cholesterol (HDL), ratio of total to HDL cholesterol, triglycerides, glycoslyated hemoglobin, body mass index, and waist circumference; 2) four inflammatory markers—interleukin-6 (IL-6), high-sensitivity C-reactive protein (CRP), soluble intercellular adhesion molecule 1 (sICAM-1), and soluble E-selectin; and 3) four neuroendocrine markers—dehydroepiandrosterone sulfate (DHEAS), cortisol, epinephrine, and norepinephrine. We also include three markers that do not represent a common biological subsystem—creatinine clearance, albumin, and homocysteine.

Self-reported health indicators. Based on self-reports from 2006, we include six health indicators: 1) global self-assessed health ("Regarding your current state of health, do you feel it is excellent, good, average, not so good, or poor?"); 2) an index of mobility limitations; 3) whether the respondent was ever diagnosed with diabetes; 4) history of cancer; 5) number of hospitalizations in the past 12 months; and 6) smoking status (never, former, current). The mobility index is based on self-reported difficulty performing each of eight physical tasks without assistance and is calculated according to the method described in Long and Pavalko [14].

Social and demographic characteristics. We control for key demographic and social characteristics that are known to be important predictors of mortality. Demographic variables comprise age, sex, urban residence, and ethnicity (Mainlander vs. Taiwanese). Social measures comprise educational attainment (in years), social integration, and perceived social support. An index of social integration is based on 10 indicators and is constructed following the strategy of Cornwell and Waite [15]. An index of perceived social support is based on four questions pertaining to potential instrumental and emotional support from family and friends. Additional details are provided in Table 1.

Measures of Discrimination

Area under the receiver operating characteristic curve (AUC). The most common approach for quantifying the predictive power of a model is the C-statistic or AUC. The receiver-operating-characteristic curve is a plot of true positive rates (sensitivity) against false positive rates (1-specificity) for all possible cutoff values that discriminate between two groups (e.g., those who died vs. those who survived). The AUC can be interpreted as the probability that the model predicts a higher probability of death for those who died than for those who survived [16]. An AUC of 0.5 indicates that the model performs no better than chance, whereas an AUC of 1.0 represents perfect accuracy.

For the purposes of evaluating the incremental value of a marker, the AUC has 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 ignores the magnitude of the difference in probabilities between models [17]; it considers the rank order of cases and noncases rather than the actual predicted probabilities [18]. In an effort to address criticisms of the AUC, researchers have developed alternative measures of discrimination based on reclassification methods.

Net Reclassification Improvement (NRI). The purpose of reclassification is to determine the extent to which inclusion of markers in a risk model improves the classification of individuals into clinically meaningful risk strata [18, 19]. The Net Reclassification Improvement (NRI) uses reclassification tables constructed separately for individuals that experience the event and those that do not [20]. It then quantifies the correct movement between risk categories (i.e., upwards for those with the event and downwards for those without 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 [19, 20]. A newer category-free version, NRI(>0), addresses this issue by redefining movement based on changes in the predicted probabilities: upward for decedents and downward for survivors [21]. 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 [1]. The NRI(>0) represents a summary measure of the correct upward versus downward movement in model-based probabilities for events and non-events [1].

The NRI(>0) is calculated as:

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

where $q_{\text{New, Event}}$ and $q_{\text{Old, Event}}$ represent the predicted probability of the event among those who experienced 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 $P\left(\text{Up}\middle|\text{Non-Event}\right)$ is the corresponding quantity among those without 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. [1] 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 [1]. Consequently, as Cook shows, one can obtain anomalous results; for example, a new model may seem inferior because of an increased number of changes in the wrong direction, even though the incorrect changes are smaller than the correct changes [22]. Pencina et al. point out that, if there is some minimum change in risk considered to be clinically meaningful, it may be preferable to calculate NRI(>*x*), where *x* represents that minimal change [1].

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 based on the difference in discrimination slopes of models with and without the new markers [20], where the discrimination slope is defined as the absolute difference in the average prediction between those who experienced the event and those who did not [23]. Thus, the IDI is calculated as:

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

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 [1].

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 [20]. 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, rewriting Equation (2) as:

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

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 [1]. Whether such dependence is desirable is debatable. Kerr et al. argue that invariance to the strength of the baseline model is not necessarily advantageous: if the baseline model is almost perfect, then the incremental value of any additional marker should be small [3]. Pencina et al. contend that the preferred metric depends on the 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 [1, 24]. 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.

Analytic Strategy

Descriptive statistics (Tables 1 and 2) are weighted to account for the sampling design and for differential response rates by various covariates. Using unweighted data, we estimate a series of Gompertz hazard models, with time measured in terms of age. Because initial tests revealed evidence of non-proportional hazards (i.e., the effect varies by age) for several covariates – perceived social support, current smoker, and the change (2000-06) in DHEAS – we include interactions between these variables and age.

In order to compare effect sizes across predictors, we standardize (mean=0, standard deviation=1) all continuous measures prior to model fitting. We transform biomarkers with a skewed distribution using a logarithm or power transformation (see Table 2) to better approximate normality, which substantially improves the model fit.

To address the questions presented above, we examine the improvement in discrimination – assessed by ΔAUC, NRI(>0), and IDI – based on a comparison of models with and without the set of indicators being evaluated. We use the coefficients from each model to compute the predicted probability of dying by the end of follow-up for each respondent. The discrimination measures are calculated from the predicted probabilities and the observed binary outcome (death vs. survival); see cph.georgetown.edu/ms410s.pdf for details. All analyses are performed using Stata 12.1 (StataCorp, College Station, TX) [25]. To compute the AUC, we use Stata's "*roctab*" and "*roccomp*" procedures.

RESULTS

Table 3 shows comparisons among selected models in terms of the three discrimination measures. Although there are no established benchmarks, Pencina and colleagues suggest ΔAUC>0.01 represents a meaningful improvement, while NRI(>0) greater than 0.6 indicates a strong contribution and NRI(>0) between 0.2 and 0.6 implies moderate improvement [1, 20]. In the results below, we use these somewhat arbitrary values as benchmarks. Researchers do not provide a corresponding gauge for IDI. However, IDI values can be interpreted as the increase in average sensitivity (given fixed specificity).

Do biomarkers retain incremental prognostic value beyond self-reports?

A comparison of Models 1 and 2 suggests that biomarkers (measured in 2006) yield substantial incremental value in predicting mortality for the period 2006-11 beyond that of self-reported health variables: ΔAUC= 0.04, NRI(>0)=0.74, and IDI=0.09.

Do changes in biomarkers yield better discrimination than one-time measurement?

A comparison of Models 2 and 3 reveals that the addition of the earlier biomarkers (2000) – i.e., incorporating change in biomarker values – yields moderate improvement: ΔAUC=0.02, NRI(>0)=0.56, and IDI=0.07.

Which cluster of biomarkers is the strongest predictor?

We evaluate the contributions of eight cardiovascular/metabolic markers (Model 4a), four inflammatory markers (Model 4b), and four neuroendocrine markers (Model 4c) by comparing each with Model 1. All three discrimination measures suggest that inflammatory markers yield more predictive power than cardiovascular/metabolic or neuroendocrine markers.

Which individual biomarkers are the strongest predictors?

Using Model 1 as the baseline, we assess the contribution of each biomarker by adding the 2006 level and 2000-06 change for that marker. Figure 1 shows the top 10 biomarkers ranked by ΔAUC, NRI(>0), and IDI.

For each discrimination measure, IL-6 is the strongest predictor and all four inflammatory markers make the top 10; sICAM-1 consistently has a high ranking as well. Although CRP is the only one of these markers used clinically, it ranks lowest of the four inflammatory markers according to all three measures.

The cardiovascular/metabolic markers in the top 10 generally rank near the bottom, and none ranks in the top 10 by all discrimination measures. Systolic blood pressure achieves a rank within the top 10 on two discrimination measures; other cardiovascular/metabolic markers appear in the top 10 only for ΔAUC (glycosylated hemoglobin) or IDI (HDL, body mass index, ratio total cholesterol to HDL).

DHEAS is the only neuroendocrine marker on all three lists. Epinephrine has a top 10 ranking for two of these, whereas norepinephrine and cortisol appear on the list for only one discrimination measure.

Among the three other unrelated markers, homocysteine ranks second or third on each list. Creatinine clearance attains the top 10 ranking for two discrimination measures, although it fails to rank higher than 8th. Serum albumin never appears in the top 10.

When we evaluate the contribution of each biomarker relative to the discrimination benchmarks described above, we find that four biomarkers yield a meaningful improvement in ΔAUC (>0.01): IL-6, DHEAS, homocysteine, and sICAM-1. One of these, IL-6, makes a strong contribution based on the NRI(>0); another six (homocysteine, sICAM-1, soluble E-selectin, DHEAS, systolic blood pressure, and CRP) yield a moderate improvement. These seven biomarkers also produce a greater than 1% improvement in sensitivity based on the IDI.

When we examine the robustness of our findings by excluding 34 respondents with CRP > 10 mg/L (indicative of acute infection), we find most of the results unchanged. In particular, the four inflammatory markers remain among the top 10 by all criteria, although they decline in rank, while homocysteine now ranks first. Three markers (homocysteine, DHEAS, and IL-6) continue to satisfy ΔAUC >0.01, and the seven markers that produce at least a moderate improvement in the NRI(>0) remain the same.

DISCUSSION

Ascertaining whether particular biomarkers enhance prediction of downstream health or mortality above and beyond conventional factors has been a contentious issue. Although there are numerous statistical criteria that need to be satisfied at an early stage of analysis (for example, statistical significance of the markers and adequate model calibration), researchers tend to focus ultimately on discrimination: does the marker improve our ability to distinguish between those who experience the event and those who do not? The merits of alternative measures of discrimination have been frequently debated, but as yet, there is no consensus about which measure is "best." Despite a large literature on evaluating novel markers, surprisingly few studies have attempted to assess the consistency of findings based on different measures of discrimination.

In this paper, we evaluate the robustness of conclusions about the utility of a set of biomarkers for predicting five-year survival in a general older population based on three frequently used discrimination measures. Several broad conclusions are consistent across the ΔAUC, NRI(>0), and IDI: (1) inclusion of biomarkers substantially enhances five-year survival prediction from a baseline model that incorporates numerous self-reported indicators of health; (2) inclusion of information on changes in biomarkers over the preceding six-year period yields a moderate improvement over one-time measurement; and (3) when considered as clusters, inflammatory markers offer stronger prediction than either cardiovascular/metabolic or neuroendocrine measures.

When we address a more specific question – which biomarkers are the strongest predictors – the findings are more nuanced. Still, there is considerable overlap in results: all three discrimination measures underscore the utility of inflammatory markers and homocysteine levels. Surprisingly, standard clinical markers that reflect lipids, obesity, and blood pressure generally have relatively low prognostic power.

At the same time, differences are apparent. For example, ΔAUC and to a lesser extent NRI(>0) favor the inclusion of neuroendocrine over cardiovascular/metabolic markers (despite only half as many variables in the former category), whereas IDI favors the cardiovascular/metabolic markers. We see this difference despite the fact that none of the discrimination measures penalizes for the number of parameters. The inconsistency stems in part from the fact that ΔAUC disproportionately weights high levels of sensitivity, where

neuroendocrine markers outperform cardiovascular/metabolic factors. At lower levels of sensitivity, the cardiovascular/metabolic factors generally perform better than the neuroendocrine markers. The rank ordering of the 10 most predictive biomarkers also varies across discrimination measures.

One important limitation of this analysis is that the benchmarks we use for the discrimination measures were originally intended for testing a *single* marker, rather than many markers. This distinction is important. For example, if an NRI(>0) of 0.60 is indicative of a strong contribution for one marker, what magnitude should be used for 19 markers? In principle, we could impose a penalty for the number of parameters added to the model. In addition, because previous research demonstrates that ΔAUC depends on the strength of the baseline model [1], we could permit the benchmark to vary accordingly, although there is little guidance on how to do so.

It is important to bear in mind that this study has not considered health outcomes beyond mortality or the many other uses of biomarkers within population-based household surveys. The value of including biomarkers undoubtedly varies across research objectives. Moreover, the cost, complex logistics and additional burden posed by the inclusion of biomarkers in large surveys of the general population must be borne in mind. Still, at least from the point of view of predicting mortality – a well-measured outcome highly correlated with myriad health measures – these findings provide strong support for biomarker collection within household surveys and moderate support for longitudinal collection of such markers – particularly with respect to inflammatory markers.

On the one hand, the consistency of most of these results across the three discrimination measures should provide some comfort to researchers. On the other hand, our findings should not signal that future evaluations based on multiple discrimination measures are superfluous. As we have shown, the degree of consistency across metrics varies with the level of detail inherent in the research question. Researchers would be wise to confirm findings with multiple metrics, for example, considering ΔAUC >0.01 as a necessary but not sufficient benchmark, and using the NRI(>0) and the IDI as robustness checks. If all three discrimination measures yield similar conclusions, the utility of the biomarker is on much firmer ground than if the metrics produce disparate results.

Upon discovering that most published research findings are inaccurate and that many of the associations reported in highly cited biomarker studies are exaggerated, Ioannidis and colleagues argue that “the standards for claiming success should be higher” [26, 27]. This analysis provides one small step in the right direction.

ACKNOWLEDGEMENTS

This work was supported by the National Institute on Aging (grant numbers R01AG16790, R01AG16661) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant number R24HD047879). Funding for the TLSA came from the Taiwan Department of Health, the Taiwan National Health Research Institute [grant number DD01-86IX-GR601S] and the Taiwan Provincial Government. SEBAS was funded by the Demography and Epidemiology Unit of the Behavioral and Social Research Program of the National Institute on Aging [grant numbers R01 AG16790, R01 AG16661]. The Bureau of Health Promotion (BHP, Department of Health, Taiwan) provided additional financial support

for SEBAS 2000. The sponsors had no involvement in the study design, data collection, analysis, interpretation of the data, writing of the report, or the decision to submit the article for publication.

We acknowledge the hard work and dedication of the staff at the Center for Population and Health Survey Research (BHP), who were instrumental in the design and implementation of the SEBAS and supervised all aspects of the fieldwork and data processing. We are also grateful to Dr. Maxine Weinstein and Dr. Germán Rodríguez for their helpful comments and suggestions regarding this manuscript.

REFERENCES

1. 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.

2. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest* 2012**; 42(2)**: 216-228. DOI: 10.1111/j.1365-2362.2011.02562.x.

3. 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.

4. Pepe MS, Janes H. Commentary: Reporting standards are needed for evaluations of risk reclassification. *Int J Epidemiol* 2011**; 40(4)**: 1106-1108. DOI: 10.1093/ije/dyr083.

5. Cornman JC, Glei DA, Goldman N, et al. Cohort profile: the Social Environment and Biomarkers of Aging Study (SEBAS) in Taiwan. [published online ahead of print September 8, 2014]. *International Journal of Epidemiology*, DOI: 10.1093/ije/dyu179.

6. Glei DA, Goldman N, Rodríguez G, Weinstein M. Beyond self-reports: changes in biomarkers as predictors of mortality. *Population and Development Review* 2014**; 40(2)**: 331-360.

7. Human Mortality Database. University of California, Berkeley (USA);Max Planck Institute for Demographic Research (Germany). www.mortality.org. Updated 2013. Accessed February 27, 2013.

8. Department of Health, Executive Yuan, R.O.C. (Taiwan). Deaths in Taiwan, 2011. http://www.doh.gov.tw/EN2006/DisplayFile.aspx?url=http%3a%2f%2fwww.doh.gov.tw%2fufile%2fdoc%2fDeaths+in+Taiwan%2c+2011-20121017.docx&name=Deaths+in+Taiwan%2c+2011-20121017.docx. Updated 2012. Accessed February 27, 2013.

9. Hoyert DL, Xu J. *Deaths: preliminary data for 2011*. Hyattsville, MD: National Center for Health Statistics; 2012; No. Vol 61, no 6.

10. Goldman N, Lin I, Weinstein M, Lin Y. Evaluating the quality of self-reports of hypertension and diabetes. *J Clin Epidemiol* 2003**; 56(2)**: 148-154. DOI: 10.1016/S0895-4356(02)00580-2.

11. Chang M, Lin H, Chuang Y, et al. *Social Environment and Biomarkers of Aging Study (SEBAS) in Taiwan, 2000 and 2006: main documentation for SEBAS longitudinal public use*

data (released 2012). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]; 2012; No. ICPSR03792-v5.

12. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res* 1999**; 8(1)**: 3-15.
13. Rubin DB. Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* 1996**; 91**: 473-489.
14. 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.
15. Cornwell EY, Waite LJ. Measuring social isolation among older adults using multiple indicators from the NSHAP study. *J Gerontol B Psychol Sci Soc Sci* 2009**; 64 Suppl 1**: i38-46. DOI: 10.1093/geronb/gbp037.
16. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004**; 23(13)**: 2109-2123. DOI: 10.1002/sim.1802.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. Cook NR. Clinically relevant measures of fit? A note of caution. *Am J Epidemiol* 2012**; 176(6)**: 488-491. DOI: 10.1093/aje/kws208.
23. Yates JF. External correspondence: decomposition of the mean probability score. *Organizational Behavior and Human Performance* 1982**; 30**: 132-156.
24. 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.
25. StataCorp. *Stata: Release 12. Statistical Software.* College Station, TX: StataCorp LP, 2011.
26. Ioannidis JP, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA* 2011**; 305(21)**: 2200-2210. DOI: 10.1001/jama.2011.713.

27. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005**; 2(8)**: E124. DOI: 10.1371/journal.pmed.0020124.

Table 1. Descriptive Statistics for Social and Demographic Characteristics, Self-reported Indicators of Health Status, and Survival Status, Weighted Analyses, Taiwan, 2006-2011, SEBAS

	Analysis sample (N=639)
Social and demographic characteristics	
Age at the 2006 exam (60-97), mean (SD)	72.0 (7.4)
Female, %	44.4
Mainlander, %	12.9
Urban resident, %	42.3
Years of completed education (0-17), mean (SD)	5.3 (4.5)
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	
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]	0.7 (1.3)
History of diabetes, %	19.9
History of cancer, %	4.8
Number of hospitalizations in the past 12 months (0-11), mean (SD)	0.3 (0.8)
Smoking status in 2006	
Never, %	59.1
Former, %	22.2
Current, %	18.7
Died between the 2006 exam and December 31, 2011, %	16.2

[a] This index was created by standardizing each of 10 indicators from the 2003 Taiwan Longitudinal Study of Aging (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) and then calculating the mean across valid items if at least eight items were valid (α=0.72). See Table S2 of cph.georgetown.edu/ms410s.pdf for more 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 [14], 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 2. Summary Statistics for Individual Biomarkers and Changes in Biomarkers, Weighted 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 (0.15)
Diastolic blood pressure (DBP)	mmHg	log	4.28 (0.15)	-0.13 (0.16)
High-density lipoprotein cholesterol (HDL)	mg/dL	log	3.84 (0.28)	-0.02 (0.22)
Ratio of total to HDL cholesterol (TC/HDL)	ratio	log	1.43 (0.27)	0.00 (0.23)
Triglycerides	mg/dL	log	4.57 (0.51)	-0.07 (0.43)
Glycosylated hemoglobin (HbA1c)	%	$-1/(\text{HbA1c})^2$	-0.03 (0.01)	0.01 (0.01)
Body Mass Index (BMI)	$\frac{\text{weight(kg)}}{(\text{height(m)})^2}$	log	3.20 (0.15)	0.00 (0.07)
Waist circumference	cm	none	84.95 (9.91)	-0.43 (5.92)
Interleukin-6 (IL-6)	pg/mL	log	1.06 (0.80)	0.23 (0.91)
C-reactive protein (CRP)	mg/L	log	-2.03 (1.13)	0.53 (1.50)
Soluble intercellular adhesion molecule 1 (sICAM-1)	ng/mL	square root	16.50 (2.92)	1.09 (2.30)
Soluble E-selectin (sE-selectin)	ng/mL	log	3.57 (0.61)	-0.18 (0.44)
Dehydroepiandrosterone sulfate (DHEAS)	µg/dL	square root	8.84 (3.22)	0.24 (2.10)
Cortisol	µg/g	log	2.68 (0.86)	-0.27 (0.96)
Epinephrine	µg/g	log	1.23 (0.58)	0.13 (0.63)
Norepinephrine	µg/g	log	3.17 (0.53)	0.20 (0.54)
Creatinine Clearance (CrCl)	ml/min	none	59.65 (20.13)	-4.98 (11.51)
Albumin	g/dL	cubed	83.65 (17.26)	-9.40 (15.31)
Homocysteine (Hcy)	µmol/L	log	2.47 (0.38)	-0.21 (0.31)

Table 3. Measures of Discrimination 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)

Model	Description	AUC	ΔAUC[c]	NRI(>0)[c]	IDI[c]
1	Baseline: Self-reported indicators only[a]	0.816			
			vs Model 1		
2	Model 1 + 19 Individual biomarkers[b]	0.857	0.041	0.74	0.09
			vs Model 2		
3	Model 2 + Changes in 19 Individual biomarkers[b]	0.877	0.020	0.56	0.07
			vs Model 1		
4a	Model 1 + 8 Cardiovascular/metabolic markers (2006 and changes 2000-06)	0.818	0.001	0.37	0.05
4b	Model 1 + 4 Inflammatory markers (2006 and changes 2000-06)	0.839	0.023	0.59	0.07
4c	Model 1 + 4 Neuroendocrine markers (2006 and changes 2000-06)	0.836	0.019	0.46	0.03

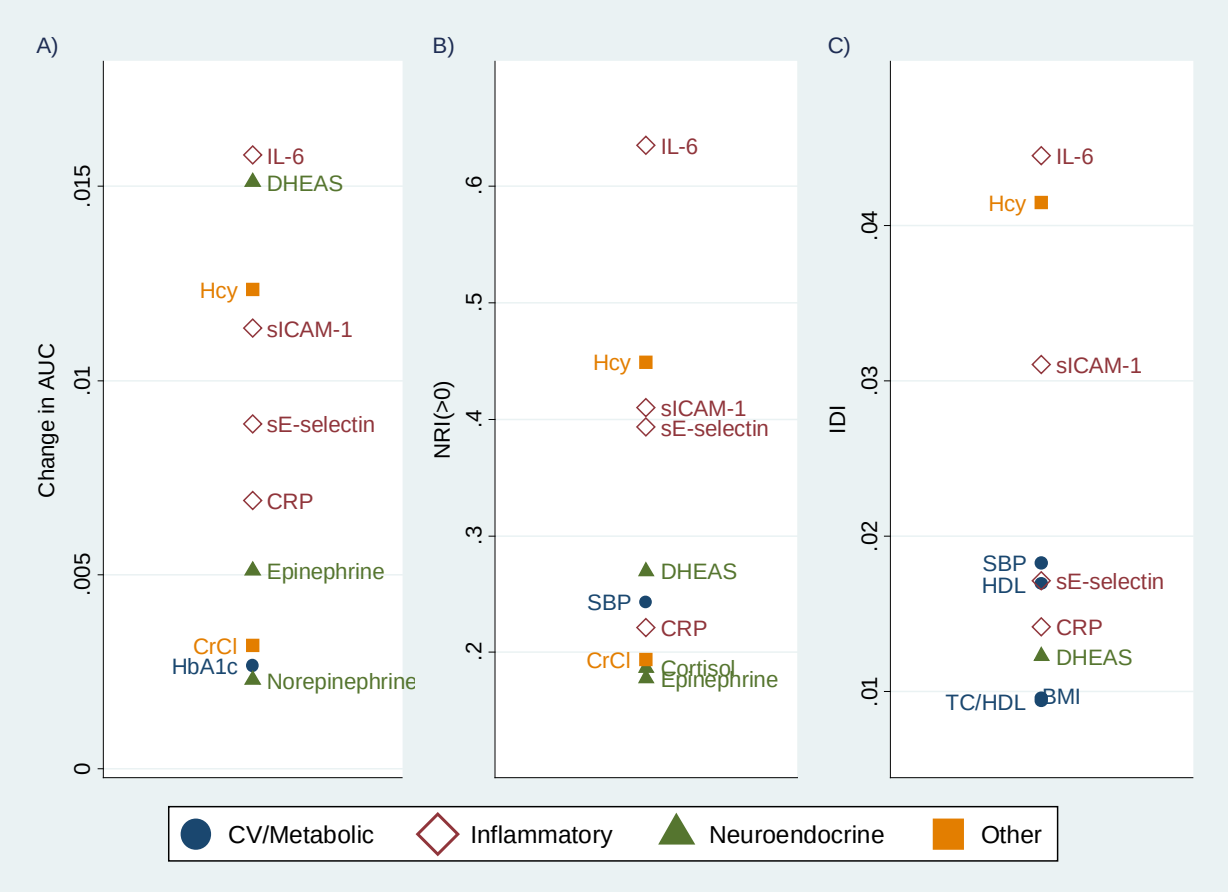
Abbreviations: AUC, area under the receiver-operating-characteristic curve; ΔAUC, change in AUC; IDI, Integrated Discrimination Improvement; NRI(>0), Continuous Net Reclassification Improvement.

[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] Measures of the improvement in discrimination (ΔAUC, NRI(>0), IDI) are based on comparisons with the model indicated. Note: Values for NRI(>0) and IDI are based on the average across five multiply imputed datasets.

Figure 1. Top 10 Biomarkers Ranked by: A) Change in AUC; B) NRI(>0); and C) IDI, Taiwan, 2000-2011, SEBAS



Measures of discrimination were calculated based on comparisons with Model 1 (Table 3), which included social and demographic characteristics and self-reported indicators of health status. For each of the 19 biomarkers, we estimated a separate model that added the 2006 level and 2000-06 change in the specified biomarker to Model 1.

Abbreviations: AUC, area under the receiver-operating-characteristic curve; BMI, Body Mass Index; CrCl, Creatinine Clearance; CRP, C-reactive protein; DHEAS, dehydroepiandrosterone sulfate; HbA1c, Glycosylated Hemoglobin; Hcy, Homocysteine; HDL, high-density lipoprotein cholesterol; IDI, Integrated Discrimination Improvement; IL-6, interleukin-6; NRI(>0), Continuous Net Reclassification Improvement; SBP, Systolic Blood Pressure; sE-selectin, soluble E-selectin; sICAM-1, soluble intercellular adhesion molecule 1; TC/HDL, ratio of total cholesterol to HDL.