American Journal of Epidemiology

Vol. 179, No. 5
DOI: 10.1093/aje/kwt298
Advance Access publication:
December 22, 2013

Practice of Epidemiology

Assessing Risk Prediction Models Using Individual Participant Data From Multiple Studies

Lisa Pennells, Stephen Kaptoge, Ian R. White, Simon G. Thompson, Angela M. Wood*, and the Emerging Risk Factors Collaboration

* Correspondence to Dr. Angela M. Wood, Strangeways Research Laboratory, Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Worts Causeway, Cambridge CB1 8RN, United Kingdom (e-mail: amw79@medschl.cam.ac.uk).

Initially submitted January 29, 2013; accepted for publication November 12, 2013.

Individual participant time-to-event data from multiple prospective epidemiologic studies enable detailed investigation into the predictive ability of risk models. Here we address the challenges in appropriately combining such information across studies. Methods are exemplified by analyses of log C-reactive protein and conventional risk factors for coronary heart disease in the Emerging Risk Factors Collaboration, a collation of individual data from multiple prospective studies with an average follow-up duration of 9.8 years (dates varied). We derive risk prediction models using Cox proportional hazards regression analysis stratified by study and obtain estimates of risk discrimination, Harrell's concordance index, and Royston's discrimination measure within each study; we then combine the estimates across studies using a weighted meta-analysis. Various weighting approaches are compared and lead us to recommend using the number of events in each study. We also discuss the calculation of measures of reclassification for multiple studies. We further show that comparison of differences in predictive ability across subgroups should be based only on within-study information and that combining measures of risk discrimination from case-control studies and prospective studies is problematic. The concordance index and discrimination measure gave qualitatively similar results throughout. While the concordance index was very heterogeneous between studies, principally because of differing age ranges, the increments in the concordance index from adding log C-reactive protein to conventional risk factors were more homogeneous.

C index; coronary heart disease; D measure; individual participant data; inverse variance; meta-analysis; risk prediction; weighting

Abbreviations: CHD, coronary heart disease; C index, concordance index; CRP, C-reactive protein; D measure, discrimination measure; NRI, Net Reclassification Index.

The derivation and assessment of risk prediction models using multiple epidemiologic studies has several advantages in comparison with analysis of single studies. These include greater precision, reduced overfitting, and increased generalizability (1, 2). Availability of individual participant data from several studies, as opposed to aggregate-level statistics, also allows detailed characterization of risk prediction models and investigation of potential effect modifiers (3). We previously described methods for investigating exposure-risk relationships using individual participant data from multiple studies (3) and proposed measures of discrimination for the stratified Cox model (4). In this paper, we extend and demonstrate methods for assessing risk prediction models using individual participant data from multiple studies based on weighted meta-analysis techniques. We describe assessment of the predictive ability of a risk prediction model, the change in predictive ability upon moving from one model to another, and the comparison of predictive abilities across different subgroups of the population. Such techniques for combining information across studies have not previously been described in detail in the literature and are of relevance to the growing number of collaborative consortia (5–13).

We illustrate these methods using data from the Emerging Risk Factors Collaboration, which comprises individual records from over 2.2 million participants in 125 prospective studies of major cardiovascular disease outcomes and cause-specific mortality in predominantly Western populations (14–17). The studies include mostly prospective cohort studies, but also some nested case-control and nested case-cohort studies. Examples presented in this paper focus on prediction models for coronary heart disease (CHD), defined as first nonfatal myocardial infarction or coronary death, and examine the predictive ability of C-reactive protein (CRP) concentration when added to conventional risk predictors. Data on CRP and conventional risk predictors at baseline were available from 37 prospective studies involving 165,856 participants without a history of cardiovascular disease, among whom 8,806 incident CHD events occurred over an average of 9.8 years of follow-up (for definitions of study names, see Web Table 1 (available at http://aje.oxfordjournals.org/)).

METHODS

Derivation of a risk prediction model over multiple studies

Initially we assume that all data are derived from prospective cohort studies; other study designs are addressed later. Risk prediction models are constructed using Cox proportional hazards models (18), stratified by study and, if applicable, by other characteristics such as sex. For studies $s = 1, \ldots, S$, with strata $k = 1, \ldots, K_s$ and individuals $i = 1, \ldots, N_s$ with baseline risk factors $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$, the probability of survival beyond time t after baseline takes the form

$$S(t|\boldsymbol{x_i}, s, k) = S_{0,s,k}(t)^{\exp(\boldsymbol{\beta x}_i)}. \quad (1)$$

The evolution of risk over time is modeled differently for each study, as represented by the nonparametric baseline survivor function $S_{0,s,k}(t)$. The vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ represents the multivariable adjusted log hazard ratios, assumed to be common to all studies, per unit increase in the risk predictors x_i. An individual's estimated linear predictor, or risk score, is simply $\hat{\boldsymbol{\beta}}\boldsymbol{x_i} = \sum_p \hat{\beta}_p x_{ip}$, and the person's absolute risk of experiencing an event by time t is estimated by $1 - \hat{S}_{0,s,k}(t)^{\exp(\hat{\boldsymbol{\beta}}\boldsymbol{x_i})}$.

Fitting the stratified model (equation 1) is a 1-stage approach to model derivation across studies (Figure 1). Alternatively, a 2-stage approach could be undertaken: First, a separate Cox proportional hazards model is fitted in each study, and then its coefficients are combined over studies to obtain $\hat{\boldsymbol{\beta}}$ using either fixed- or random-effects (multivariate) meta-analysis (3, 19, 20). A 2-stage random-effects meta-analysis has the advantage of allowing for heterogeneity in the true coefficients between studies, giving a larger variance for $\hat{\boldsymbol{\beta}}$. A multivariate meta-analysis combines estimates for the vector of correlated coefficients, taking account of its covariance matrix, over the multiple studies; separate univariate meta-analyses ignore the correlations between the coefficients. With the 2-stage approach, additional estimation is required to obtain study-specific baseline survivor functions necessary for making absolute risk predictions. The 1- and 2-stage approaches often give similar $\hat{\boldsymbol{\beta}}$ estimates (21), and

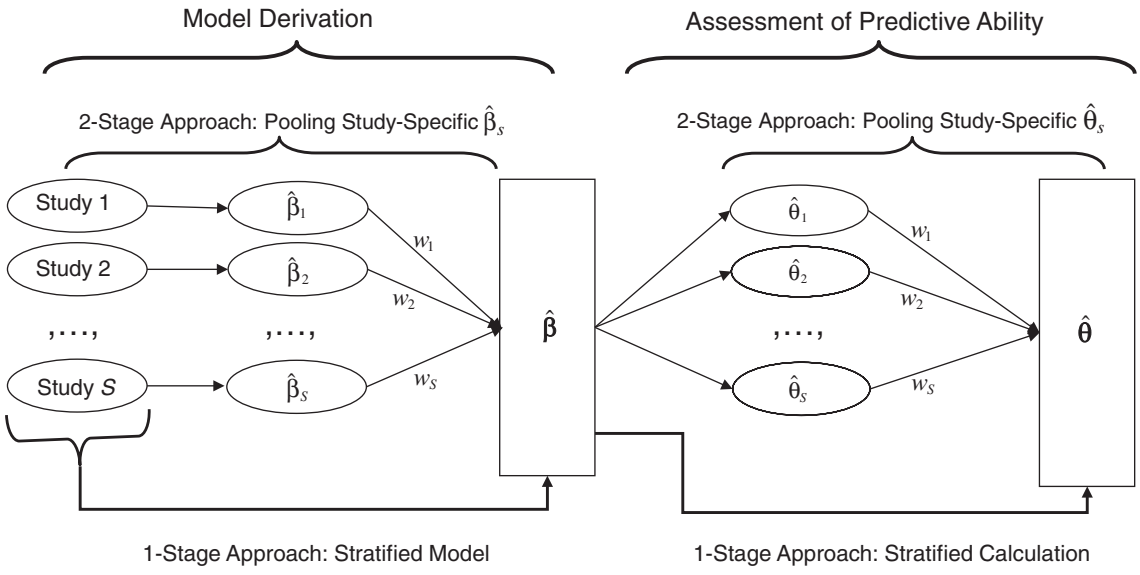

Figure 1. Overall schemes for model derivation and testing of predictive ability over multiple studies. In the model derivation process, study-specific data sets are used to estimate the pooled vector of coefficients ($\hat{\boldsymbol{\beta}}$) for the included risk predictors, either by means of a 1-stage stratified model or by a 2-stage approach applying meta-analysis of study-specific estimates. In assessment of predictive ability, the pooled $\hat{\boldsymbol{\beta}}$ is used to calculate the pooled discrimination statistic, either using a 1-stage stratified approach or by meta-analyzing study-specific estimates in a 2-stage approach. w_S represents study-specific weights applied in meta-analysis approaches; possible choices are described in the text.

hence risk scores, and the 1-stage model (equation 1) has the advantage of simplicity.

The selection of risk predictors may depend on several factors, including statistical significance, clinical importance, costs, and predictive ability. This paper focuses on the latter. The primary descriptions in this paper assume that the time scale used is duration of time in the study and that the proportional hazards assumption is met (3, 22). Assessment of predictive ability given more complex model formulations is discussed later.

Example: Deriving risk prediction models using data from the Emerging Risk Factors Collaboration. Examples presented in this paper are CHD risk models with conventional risk predictors and log CRP. Deaths from other causes or other nonfatal vascular outcomes (e.g., stroke) are regarded as censored observations. There is considerable variation in the censoring proportions (which equal 1 minus the event proportions) across studies (Web Figure 1). Table 1 shows summary statistics and $\hat{\boldsymbol{\beta}}$ using each of the described approaches. The 1-stage stratified model (equation 1) and the 2-stage fixed- and random-effects approaches all yield similar values for $\hat{\boldsymbol{\beta}}$; the standard errors for the random-effects method are greater, reflecting between-study heterogeneity. We checked the proportional hazards assumption by assessing the interaction between log CRP and time in a time-dependent Cox model using a 2-stage approach (i.e., study-specific interactions were first calculated and then combined using random-effects meta-analysis (3)). There was no evidence against the proportional hazards assumption (17), and the 1-stage prediction model (equation 1) is used for all further examples. The 2 corresponding risk scores (without and with log CRP) are given in the footnotes of Table 1. The distributions of the linear predictors are approximately normally distributed (Web Table 2).

Measures of discrimination

Measures of discrimination quantify the degree to which a model can predict the order of events. Two such measures are the concordance index (C index) (22, 23) and the discrimination measure (D measure) (24), which are pertinent because of their relevant interpretation, familiarity for the intended clinical and epidemiologic audience, and low sensitivity to censoring in the absence of marked skewness of the linear predictor. In a single unstratified study, the C index estimates the probability of concordance between predicted risk and the observed order of events for a randomly selected pair of participants (22, 23). Only informative pairs (where it is possible to determine which person suffered the first event) are used: This introduces some sensitivity to censoring (25), which we ignore because currently available solutions either assume correct model specification (25) or do not accommodate censoring by the end of follow-up (26). The D measure estimates the mean log hazard ratio for the event of interest for a randomly selected pair of participants (for one individual in the top half of the predicted risk distribution versus another individual in the bottom half) (24). The variance of the C index can be calculated by bootstrapping or by means of a jackknife

Table 1. Characteristics of Study Participants in the Emerging Risk Factors Collaboration and Comparison of Log Hazard Ratios for Coronary Heart Disease in Multivariable-Adjusted Models

	Mean (SD)	No. of Subjects	%	Multivariable-Adjusted Log HR[a] (SE)			Heterogeneity	
				1-Stage Stratified Model[b,c]	2-Stage Fixed-Effects Model[d]	2-Stage Random-Effects Model[d]	I^2	95% CI
Age at survey, years	64.2 (8.6)			0.567 (0.013)	0.565 (0.013)	0.529 (0.043)	76	67, 82
Male sex		81,732	49	NA	NA	NA	NA	NA
Current smoking[e]		35,577	21	0.516 (0.024)	0.529 (0.024)	0.515 (0.050)	63	48, 73
Systolic blood pressure, mm Hg	131 (19)			0.202 (0.009)	0.203 (0.009)	0.211 (0.017)	30	0, 53
History of diabetes[e]		10,790	7	0.557 (0.038)	0.587 (0.037)	0.600 (0.049)	24	0, 49
Total cholesterol, mmol/L	5.84 (1.06)			0.234 (0.010)	0.235 (0.010)	0.216 (0.018)	32	0, 54
HDL cholesterol, mmol/L	1.27 (0.38)			−0.247 (0.014)	−0.240 (0.014)	−0.232 (0.023)	52	32, 67
Log CRP, mg/L	0.55 (1.09)			0.206 (0.012)	0.207 (0.012)	0.201 (0.013)	9	0, 38

Abbreviations: CI, confidence interval; CRP, C-reactive protein; HDL, high-density lipoprotein; HR, hazard ratio; NA, not applicable; SD, standard deviation; SE, standard error.

[a] Log HR for coronary heart disease per 1-SD increase or in comparison with the relevant reference category, using data from 37 studies (165,856 participants with 8,806 cases of coronary heart disease).

[b] Implies that log HRs were estimated using the stratified model described in equation 1.

[c] The 1-stage stratified model was used to construct the following risk scores (note that log HRs now represent a 1-unit increase in continuous risk factors)—risk score without CRP: $0.068 \times \text{age} + 0.576 \times \text{smoker} + 0.012 \times \text{systolic blood pressure} + 0.584 \times \text{diabetic} + 0.221 \times \text{total cholesterol} - 0.756 \times \text{HDL cholesterol}$; risk score with CRP: $0.066 \times \text{age} + 0.516 \times \text{smoker} + 0.011 \times \text{systolic blood pressure} + 0.557 \times \text{diabetic} + 0.220 \times \text{total cholesterol} - 0.652 \times \text{HDL cholesterol} + 0.189 \times \log \text{CRP}$.

[d] Implies that log HRs were estimated by meta-analyzing study-specific estimates assuming fixed or random effects, respectively.

[e] Reference categories were non–current smoker for smoking and nondiabetic for history of diabetes.

procedure (27), whereas the variance of the D measure is simply a log hazard ratio variance. When stratification (e.g., by study and sex) is used, the selection of pairs is constrained to be within the same stratum.

Calculation of measures of discrimination using multiple studies

A 2-stage approach can be used to estimate discrimination measures over multiple studies (Figure 1). Firstly, the discrimination measure is estimated within each study s, denoted by $\hat{\theta}_s$, with corresponding standard error (SE) $\hat{\sigma}_s^2$. The study-specific estimates are then combined using a weighted average to obtain the pooled estimate $\hat{\theta}$:

$$\hat{\theta} = \frac{\sum w_s \hat{\theta}_s}{\sum w_s} \tag{2}$$

$$\text{with SE } \hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{\sum w_s^2 \hat{\sigma}_s^2}{(\sum w_s)^2}}, \tag{3}$$

where w_s is the weight applied to each study's estimate. The estimated $\hat{\theta}$ represents the discrimination measure when pairs of individuals are selected via a 2-stage sampling scheme: First one of the S studies is selected with probability proportional to w_s, and then a pair is selected at random from that study. We choose w_s as the study-specific number of events, since this is the principal determinant of study precision. Since this weighting scheme is concerned with sampling from existing studies only, it is not relevant to allow for heterogeneity in $\hat{\theta}_s$ across studies.

Possible alternative weights include inverse-variance weights from a fixed- or random-effects meta-analysis. The latter results in a C index which estimates the probability of concordance for a randomly selected pair of participants from a new study, sampled from a distribution from which the existing studies are believed to have come.

Alternatively, 1-stage stratified calculations of the discrimination measure could be used, stratifying by study (4) (Figure 1). For the stratified D measure, this gives study weights similar to the number of events. For the stratified C index, however, studies receive weights according to the number of contributing informative pairs, which generally depends on the total number of study participants. As a result, large studies with few events can receive substantial weight, which may be unappealing.

The impact of heterogeneity on the imprecision of the pooled estimate of discrimination can be quantified by the I^2 statistic, defined as the percentage of variance in the study-specific point estimates that is attributable to true between-study heterogeneity as opposed to sampling variation (28). Values of I^2 close to 0% correspond to lack of heterogeneity, and values close to 100% correspond to heterogeneity much larger than the sampling variation. The primary determinants of heterogeneity in study-specific estimates of discrimination $\hat{\theta}_s$ are: 1) study-specific distributions of the risk predictors, with wider ranges of continuous risk predictors leading to higher values of $\hat{\theta}_s$, and 2) variation in the relevance of the pooled $\hat{\boldsymbol{\beta}}$ to individual studies.

Example: Calculation of C index and D measure. Web Figure 1 illustrates study-specific C indices for a conventional CHD prediction model (including all predictors in Table 1 except log CRP), with pooled estimates derived using various weighting schemes. The second and third columns show that the proportion of events varies from 0.5% to 20%. Weighting by the number of informative pairs gives inappropriate weighting across studies, with 2 large studies (the Reykjavik Study and the Women's Health Study) receiving 57% of the combined weight. When weighting is done by number of events or by inverse variance assuming fixed effects, large studies with few events are assigned comparatively less weight, but the contribution of studies with many events, such as the Reykjavik Study, remains substantial. In contrast, weights assuming random effects are more uniformly assigned across studies. This is expected in the presence of large between-study heterogeneity, as is the noticeably wider 95% confidence interval for the pooled estimate. This latter approach allows calculation of a 95% prediction interval (29) to indicate the range of values that might be expected in a new study when there is between-study heterogeneity (Web Figure 1).

Similar results are seen with the D measure (Web Figure 2), and there is strong correlation between the study-specific C index and D measure (Web Figure 3). Heterogeneity in study-specific absolute values for both measures is substantial (I^2 = 93% for the C index and I^2 = 91% for the D measure). Meta-regression (21, 30) reveals strong positive correlations between study-specific $\hat{\theta}_s$ (C index or D measure) and the standard deviation of age (Figure 2, top panels). After the meta-regression adjustment, the I^2 value for the C index is reduced to 75%, remaining substantial.

Meta-regression also reveals correlations between study-specific $\hat{\theta}_s$ and the standard deviation of the prognostic index and, to a lesser extent, its skewness and kurtosis (Web Figure 4, top panels), probably due to the extensive censoring. Calculations are stratified by sex, preventing sex from contributing to calculation of $\hat{\theta}_s$ and eliminating between-study heterogeneity caused by different proportions of males and females. I^2 values in Table 1 indicate moderate heterogeneity in study-specific $\hat{\boldsymbol{\beta}}_s$ for some predictors (particularly age and smoking), which may also explain some of the heterogeneity in $\hat{\theta}_s$.

Calculating a change in discrimination using multiple studies

Often interest is in the difference in $\hat{\theta}$ between 2 alternative models, denoted $\hat{\Delta} = \hat{\theta}_{\text{model 2}} - \hat{\theta}_{\text{model 1}}$. As before, we combine study-specific differences in the C index or D measure, $\hat{\Delta}_s = \hat{\theta}_{s,\text{model 2}} - \hat{\theta}_{s,\text{model 1}}$, with variances $\hat{\sigma}^2_{\hat{\Delta}s}$:

$$\hat{\Delta} = \frac{\sum w_s \hat{\Delta}_s}{\sum w_s} \tag{4}$$

$$\text{with SE } \hat{\sigma}_{\hat{\Delta}} = \sqrt{\frac{\sum w_s^2 \hat{\sigma}^2_{\hat{\Delta}s}}{(\sum w_s)^2}}, \tag{5}$$

where $\hat{\sigma}^2_{\hat{\Delta}s}$ for the C index difference is directly estimable using the jackknife procedure (27) and for the D measure the difference is obtained using nonparametric bootstrapping.

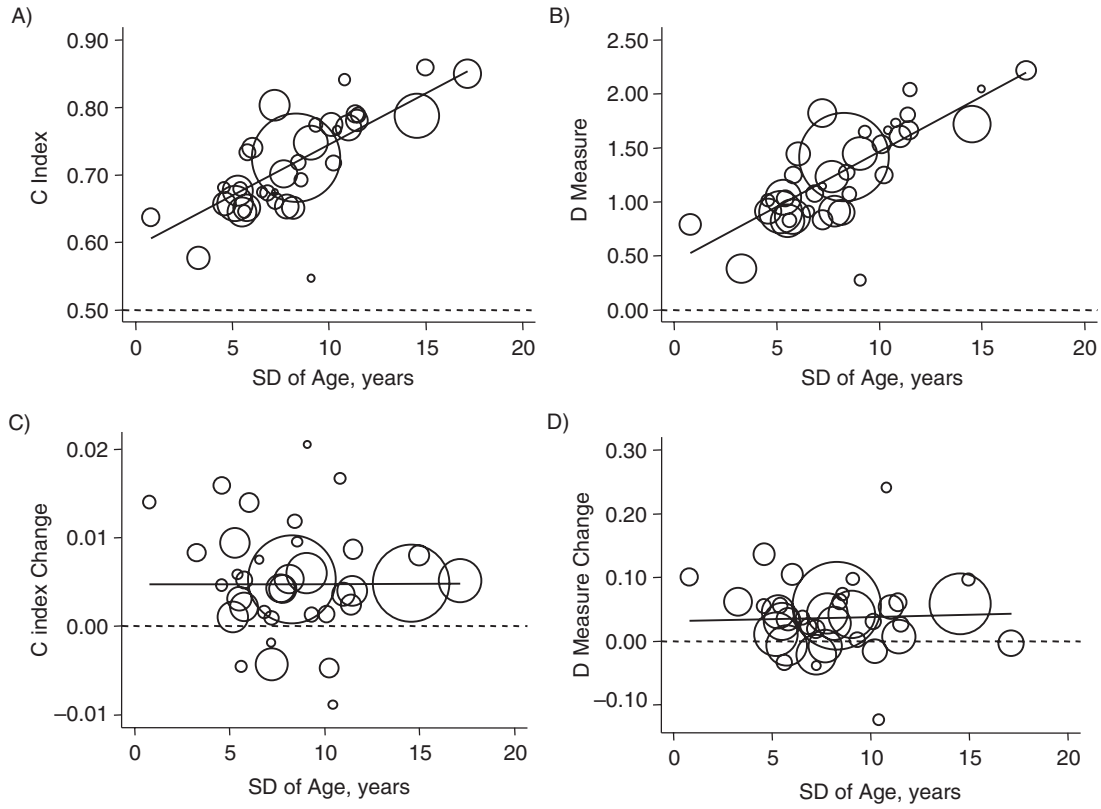

Figure 2. Meta-regression of study-specific concordance index (C index) and discrimination measure (D measure) for model 1, and subsequent changes upon addition of log C-reactive protein, on the study-specific standard deviation (SD) of age. Model 1 included conventional risk factors: age, smoking status, systolic blood pressure, history of diabetes, total cholesterol, and high-density lipoprotein cholesterol, and results are stratified by sex. The size of each circle represents the inverse variance weight applied to each study in the meta-regression.

When pooling study-specific changes, weighting by the number of events is attractive for the same reasons as those discussed above for pooling absolute values. This scheme also ensures consistency between the difference in the pooled model-specific $\hat{\theta}$ estimates $(\hat{\theta}_{\text{model 2}} - \hat{\theta}_{\text{model 1}})$ and the result obtained by pooling the within-study differences $\hat{\Delta}_s$. Such consistency is not true of inverse-variance weighting approaches, which may give a pooled estimate close to the null, since small values of $\hat{\Delta}_s$ tend to have small variances.

When model 2 extends model 1 through added predictors, $\hat{\Delta}_s$ represents the incremental predictive ability of the added predictors, and heterogeneity in $\hat{\Delta}_s$ depends on 1) the study-specific distributions of the added predictors (wider ranges leading to greater $\hat{\Delta}_s$) and 2) the relevance of the pooled $\hat{\boldsymbol{\beta}}$ for the added predictors to individual studies. In addition, since the C index has an upper bound of 1, improvements in the C index are more difficult to achieve for higher starting values. Given heterogeneity in study-specific C indices for model 1, we might expect consequent heterogeneity in $\hat{\Delta}_s$. Since the D measure is a log hazard ratio, this potential "ceiling effect" does not apply.

Example: Calculating a change in discrimination. For our example, model 1 contains conventional CHD risk predictors nested within model 2, which additionally contains log CRP (Table 1). There are significant increases in the C index and D measure upon the addition of log CRP under all weighting schemes (Web Figures 5 and 6). The clinical interpretation of this is discussed in detail elsewhere (17). Heterogeneity in $\hat{\Delta}_s$ is less than that for absolute values $\hat{\theta}_s$ ($I^2 = 0\%$ and $I^2 = 26\%$ for C-index and D-measure changes, respectively). This lack of heterogeneity can be attributed to similarity in the distribution of log CRP across studies and to homogeneity in $\hat{\beta}_s$ for this predictor ($I^2 = 9\%$). $\hat{\Delta}_s$ is also independent of study-specific age range (Figure 2, bottom panels), as well as the standard deviation, skewness, and kurtosis of the prognostic index (Web Figure 4, bottom panels), and we see little impact of the ceiling effect with the C index. The correlation between within-study changes in the C index and D measure is strong (Web Figure 3).

Subgroup-specific measures of discrimination

Of possible interest are subgroup-specific changes in discrimination, $\Delta_m = \theta_{m,\text{model 2}} - \theta_{m,\text{model 1}}$ for $m = 1, \ldots, M$ subgroups, upon addition of a new predictor. These can be estimated as follows: 1) using equation 1, fit model 1 (without the new predictor) and model 2 (with the new predictor and its interaction with the subgroup variable); 2) calculate study- and subgroup-specific discrimination measures for each model as

$\hat{\theta}_{s,m,\text{model 1}}$ and $\hat{\theta}_{s,m,\text{model 2}}$ and their difference, $\hat{\Delta}_{s,m}$; and 3) pool $\hat{\Delta}_{s,m}$ and their variance estimates across studies as in equations 4 and 5 using study weights equal to the number of events within the subgroup, to obtain subgroup-specific estimates $\hat{\Delta}_m$ and corresponding standard errors $\hat{\sigma}_{\hat{\Delta}_m}$. The null hypothesis that Δ_m is the same across all subgroups can be tested using a χ^2 test with $M-1$ degrees of freedom. To maintain within-study comparisons between subgroups, only studies with data on all subgroup levels are used (e.g., only studies with both men and women are used to compare sex-specific subgroups). This avoids erroneous conclusions resulting from between-study comparisons (3).

Example: Calculating measures of discrimination across subgroups. Figure 3 shows that log CRP appears to provide less improvement in discrimination among women and nonsmokers, although these results will require confirmation elsewhere before being adopted into clinical guidelines (17).

Other issues

Time-dependent risk predictions. Measures of discrimination assess how well participants are ranked in terms of risk predictions. In a proportional hazards model, where duration of time in the study is employed as the time scale and baseline covariates are used, ranking of t-year risk does not change with t; hence, measures of discrimination are stable over time and $\hat{\boldsymbol{\beta}}\boldsymbol{x_i}$ is sufficient for their calculation. If time-dependent covariates are introduced or if nonproportional hazards are modeled, the ranking of t-year risk can change with t. In such situations, we suggest calculating the predicted risk estimates for each individual for a single (or a selection of) fixed time point(s) t. Each set of t-year risk predictions can be used to rank individuals and calculate t-year-specific measures of discrimination using the methods previously described, but with censoring of follow-up at the selected t so that only the order of events occurring before t is considered. Since measures of discrimination will now change with time, careful choice of t is required. If a selection of fixed points t is used, then it may be useful to plot the t-year-specific measures against t.

Similar considerations are relevant when using age as the time scale (31–33); in this case, participant entry into the model is staggered (with entry at starting age) and, again, ranking of t-year risk changes with choice of t. Here, risk

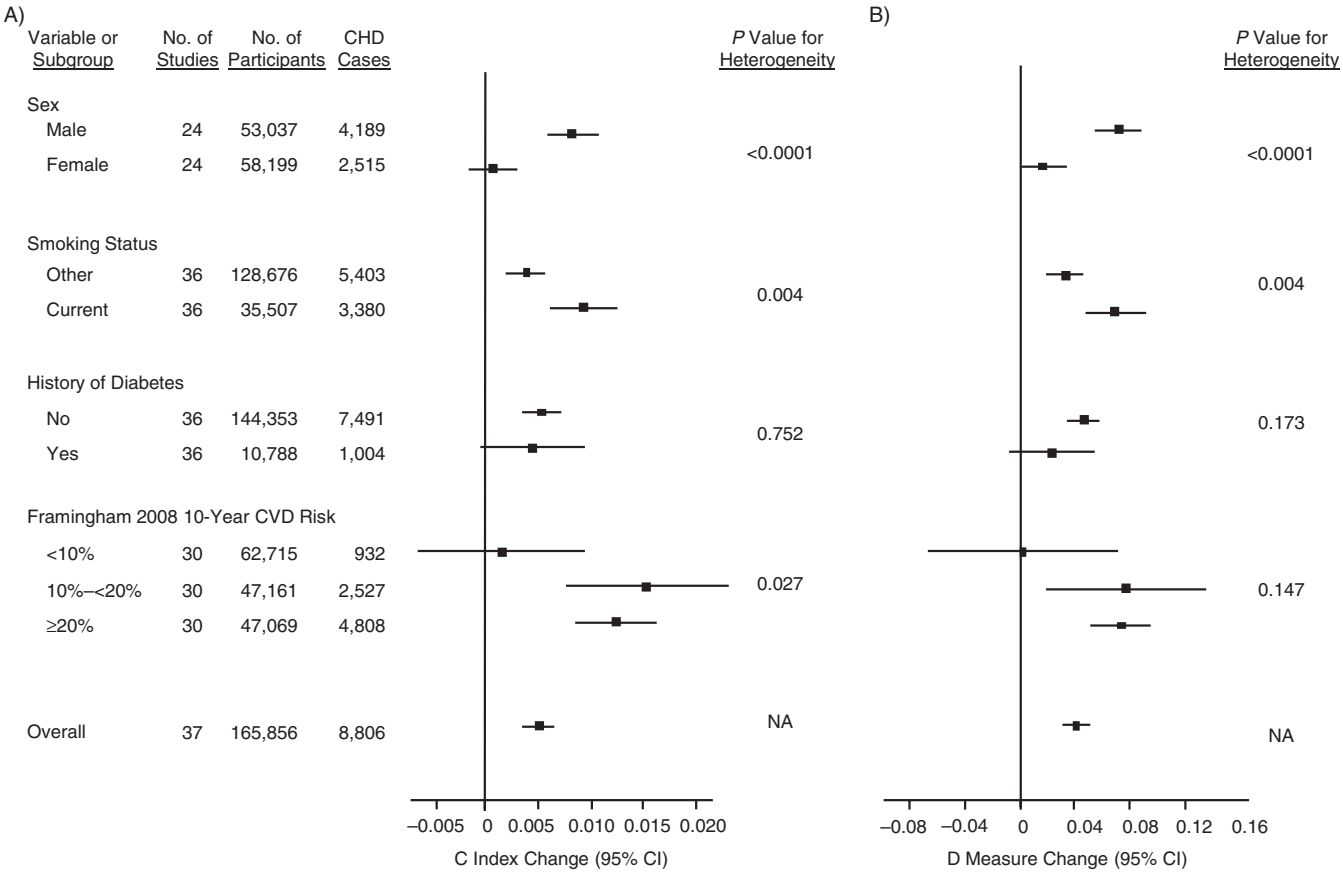

Figure 3. Changes in the concordance index (C index) (section A) and the discrimination measure (D measure) (section B) upon movement from model 1 to model 2 within various population subgroups. Model 1 included conventional risk factors: age, smoking status, systolic blood pressure, history of diabetes, total cholesterol, and high-density lipoprotein cholesterol, and results are stratified by sex. Model 2 additionally included log C-reactive protein and an interaction term for interaction between this predictor and each subgroup factor. Bars, 95% confidence intervals (CIs). CHD, coronary heart disease; CVD, cardiovascular disease; NA, not applicable.

from "age-at-entry" to "age-at-entry plus *t* years" can be used to rank participants in calculation of the discrimination measure. With this approach, follow-up should be censored at "age-at-entry plus *t* years." A simpler approach is possible for the D measure, in which the original algorithm is used but with age as the time scale in the Cox models. This will yield lower D measures, since age is effectively adjusted for in the discrimination calculation.

Case-control studies. It is possible to estimate predictive ability in case-control studies using the area under the receiver operating characteristic curve (22). However, matched case-control studies create 2 problems: 1) coefficients for matched variables are essentially meaningless (or null), leading to distortion of risk predictions, and 2) the restricted distribution of matched variables means that discrimination appears reduced. Hence, as has been reported previously (34, 35), values of discrimination obtained are commonly inconsistent with those from cohort studies.

Example: C statistic for nested case-control studies. Figure 4 illustrates that the *C* statistic is substantially lower in nested age-matched case-control studies than in cohort studies, and the corresponding change upon addition of log CRP is greater.

Measures of reclassification

Measures of reclassification quantify the extent to which individuals are more appropriately classified into risk categories using a new model versus an old model. Participants are placed into predefined risk categories based on their predicted absolute risk of experiencing an event by time *t* according to each model. Reclassification can be quantified using the Net Reclassification Index (NRI) (36), which is the sum of 2 proportions: 1) the proportion of events by time *t* that move up through the risk categories upon using the new model and 2) the proportion of nonevents at time *t* that move down through the risk categories upon using the new model. We suggest reporting these 2 meaningful proportions (as an "event NRI" and "nonevent NRI," respectively) along with an overall NRI. Participants censored before *t* years are excluded from these calculations.

For multiple studies, and having derived the prediction model using only studies with at least *t* years of follow-up, a 1-stage approach for the calculation of the NRI across multiple studies can be applied by calculating the 2 proportions across all studies. A 2-stage approach could also be taken, in which the NRI is calculated within each study before pooling. However, study-specific estimates of the NRI can be unstable if few participants experience an event and, hence, very few events change categories. It is also unclear which weights to apply; while weighting by the number of events is intuitively sensible for measures of discrimination and for the "event NRI" component, it is less relevant for the "nonevent NRI." Weighting the "event NRI" and the "nonevent NRI" by the number of events and nonevents, respectively, is equivalent to the 1-stage approach. Inverse-variance weighting produces results closer to the null, since studies with few movements between risk categories will have small standard errors and

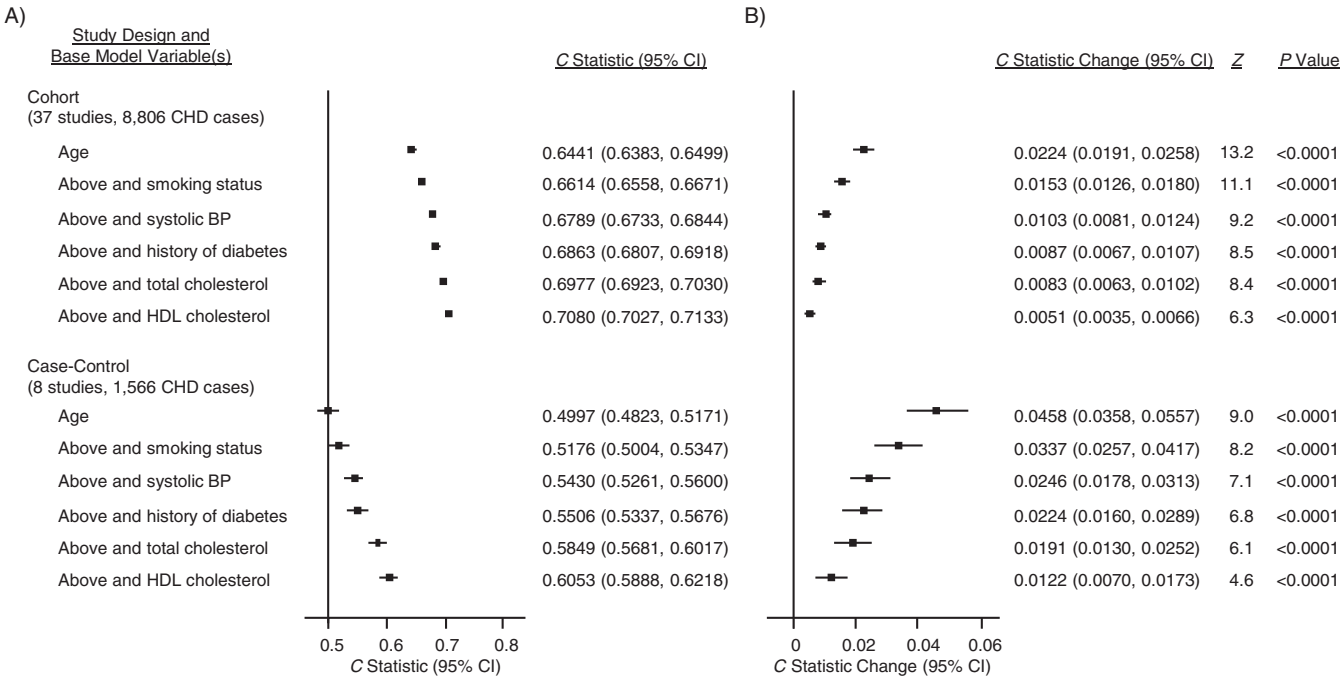

Figure 4. Comparison of *C* statistics for the cohort and case-control study designs. The concordance index (C index) is shown for cohort studies, and the area under the receiver operating characteristic curve is shown for case-control studies. Section A shows values for base models with the progressive addition of conventional risk factors for coronary heart disease (CHD), and section B shows the resulting change in the *C* statistic upon addition of log C-reactive protein (CRP) to each base model. Bars, 95% confidence intervals (CIs). BP, blood pressure; HDL, high-density lipoprotein.

receive greater weight. The most appropriate weighting scheme to use may depend on the similarity of studies in terms of risk distribution, which affects the proximity of participant risk predictions to the risk category boundaries and hence the degree of movement between categories. Ideally the risk distribution within each study should match that of the target population, which would make inverse-variance weighting more clinically relevant. In the absence of data of this type, reweighting calculations to mimic movement that would be expected in a standard target population (e.g., a standard European population) is a possibility.

Calculation of the NRI is less feasible for case-control studies, since it is not possible to directly derive absolute risk estimates from these studies. Matching (particularly by age) also affects the proportions of participants placed into each category such that they are not representative of the target population.

Example: Calculation of the NRI. Figures 5 and 6 illustrate different weighting schemes in the estimation of the NRI for models with and without log CRP using conventional 10-year risk categories (<10%, 10% to <20%, and ≥20%). Studies with fewer than 10 years of follow-up and participants censored before 10 years are excluded, and thus the NRI estimates are not directly comparable with the C index or D measure. The power of the NRI is also lower, since it relies on only a few categories of predicted risk. Inverse-variance weighting (particularly fixed effects) gives large weight to 1 low-risk study (the Women's Health Study) in which the majority of participants (all female) are nonevents in the lowest risk category. Since there is little movement between risk categories, the study's NRI estimate has a small standard error and receives a large inverse-variance weight. Our current recommendation is to use the 1-stage calculation.

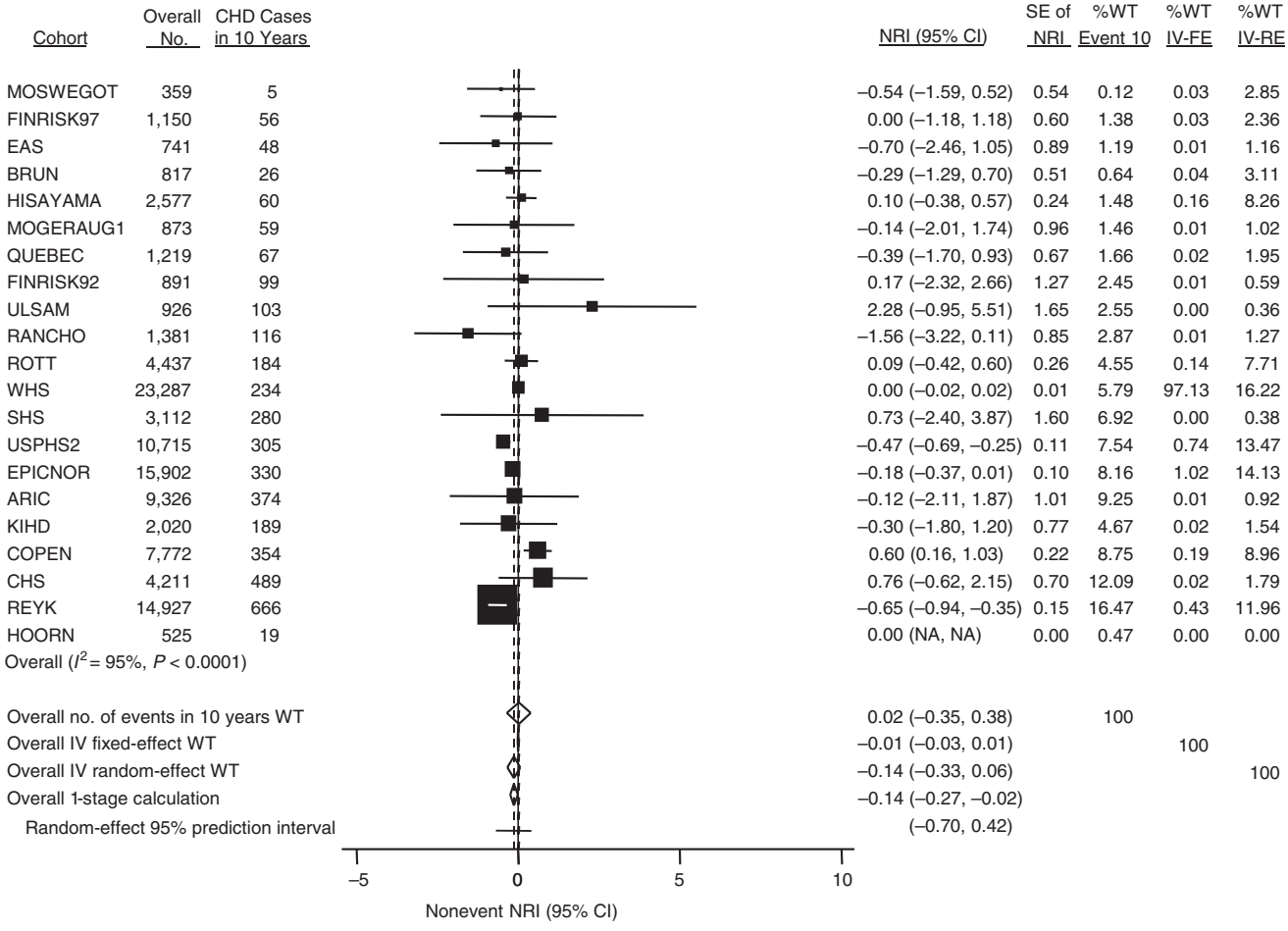

Figure 5. Study-specific estimates of nonevent Net Reclassification Index (NRI) upon application of model 2 versus model 1 and overall estimates obtained using a 1-stage approach and by meta-analysis using 3 alternative weighting schemes in the Emerging Risk Factors Collaboration. Model 1 included conventional risk factors: age, smoking status, systolic blood pressure, history of diabetes, total cholesterol, and high-density lipoprotein cholesterol, and results are stratified by sex. Model 2 additionally included log C-reactive protein. The 3 weighting schemes illustrated are 1) number of contributing events occurring before 10 years (Event 10), 2) inverse-variance weights assuming fixed effects (IV-FE), and 3) inverse-variance weights assuming random effects (IV-RE). There was no reclassification observed among nonevents in the Hoorn Study (shown at the bottom), and therefore it does not contribute to the inverse-variance-weighted pooled estimates due to undefined weight. Bars, 95% confidence intervals (CIs). CHD, coronary heart disease; NA, not applicable; SE, standard error; WT, weight. Definitions of study names are given in Web Table 1.

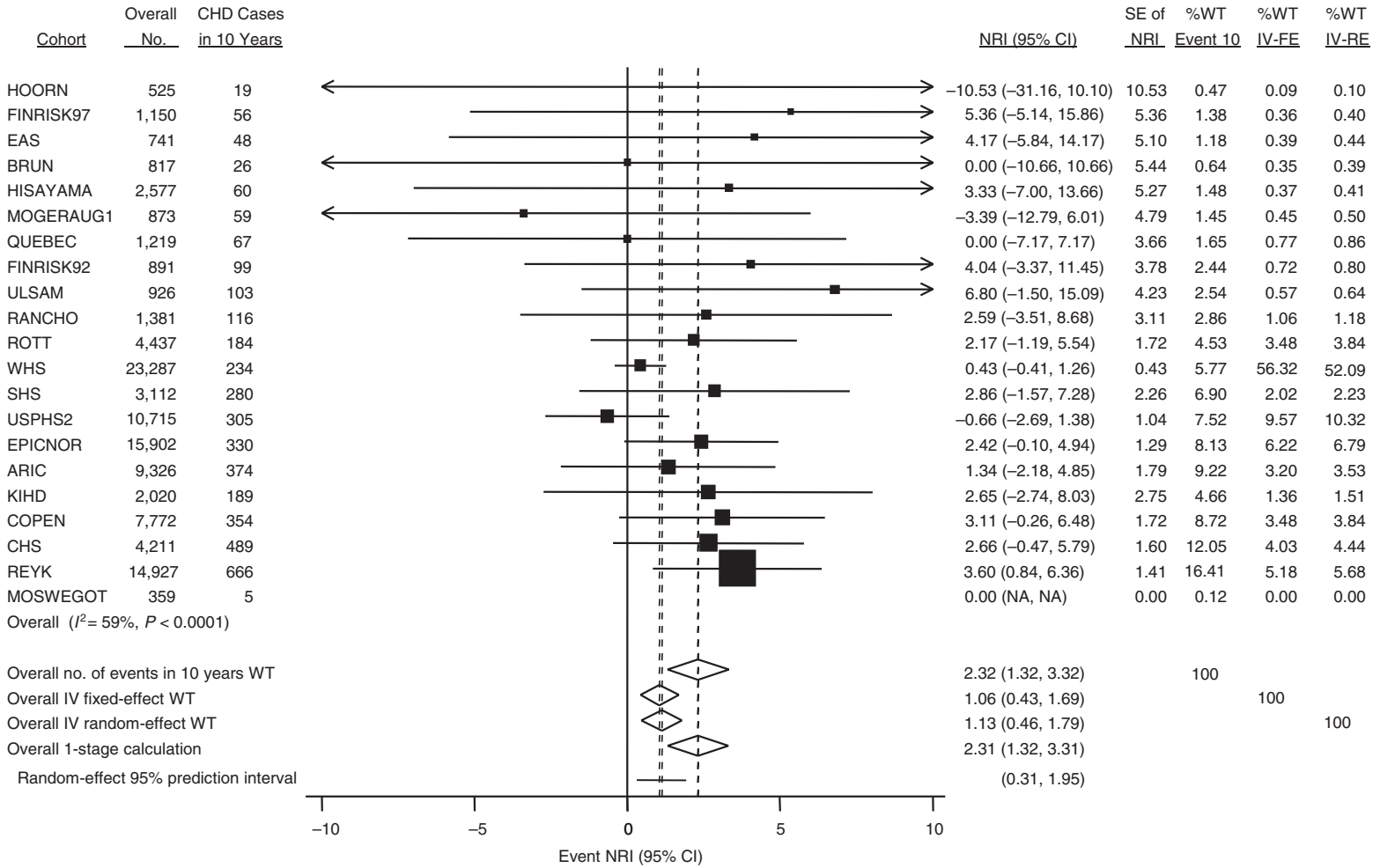

Figure 6. Study-specific estimates of event Net Reclassification Index (NRI) upon application of model 2 versus model 1 and overall estimates obtained using a 1-stage approach and by meta-analysis using 3 alternative weighting schemes in the Emerging Risk Factors Collaboration. See the legend of Figure 5 for explanations. There was no reclassification observed among events in the MONICA Göteborg Study (shown at the bottom), and therefore it does not contribute to the inverse-variance-weighted pooled estimates due to undefined weight. Bars, 95% confidence intervals (CIs). CHD, coronary heart disease; NA, not applicable; SE, standard error; WT, weight. Definitions of study names are given in Web Table 1.

The prospective NRI. The prospective NRI (37) allows inclusion of participants censored before t years and requires estimating (using Kaplan-Meier methods) the probability of having an event among 1) all participants, 2) those who move up through the risk categories, and 3) those who move down through the risk categories. Its standard error can be obtained by bootstrapping. The prospective NRI is generally more stable within studies (since censored participants are included, increasing numbers), and its calculation lends itself better to within-study calculation and the 2-stage approach using equations 2 and 3.

DISCUSSION

In this paper, we have demonstrated methods for combining measures of predictive ability across multiple studies. Relevant Stata code (StataCorp LP, College Station, Texas) is available from the University of Cambridge (http://www.phpc.cam.ac.uk/ceu/research/erfc/stata). We have described assessment of the predictive ability of a risk prediction model, the change in predictive ability upon moving from one model to another, and the comparison of predictive abilities across subgroups of the population. Various approaches to weighting estimates of predictive ability when pooling across studies have been discussed, and we recommend using the number of events in each study. Study designs other than the prospective cohort study and risk prediction models other than the Cox proportional hazards model with duration of time in the study as the time scale have also been considered. The clinical implications of using CRP concentration as an additional predictor of CHD risk are considered in detail elsewhere (17).

We used the C index (22), the D measure (24), and measures of reclassification (36) to illustrate the additional predictive ability of CRP in prediction of 10-year CHD risk. For these data, the C index and D measure led to similar conclusions with similar statistical power, whereas reclassification measures were not comparable because of their different interpretations, use of study-specific absolute risks instead of linear predictors, use of risk cutoffs, and exclusion of censored observations. Calculation of the D measure and its standard error required the least computational time.

Study-specific estimates of discrimination were dependent on the risk predictors' distributions. The implication is that any pooled estimate of discrimination represents a value applicable to a population with "average" risk predictor ranges. Caution should be applied when comparing estimates of discrimination across studies with large differences in the risk predictors' distributions. In our examples, the changes in discrimination were less heterogeneous and therefore more reliably combined across studies.

Other approaches with which to assess prediction models exist. These include measures of explained variation, which quantify the proportion of variation in the outcome that can be explained by the predictors in the model. Few such measures appear to adequately deal with censored data, and these approaches have proven difficult to adapt to the multistudy context (4). Others, such as the R^2_D extension to the D measure (38), could be applied. Measures of calibration generally compare observed and predicted risks within groups (e.g., deciles) and quantify any evidence for lack of model fit with a P value (39, 40). Calibration is important for the assessment of a new proposed model for a target population, but it is not the main issue when comparing the predictive ability of alternative models. We have also not considered validation approaches (41). Internal validation has not been necessary, because the overall data sets used are of substantial size and overfitting is minimal (42). External validation is more relevant when a new risk score is proposed and its generalizability is of interest (1). Addressing overfitting is more important when combining measures of discrimination using inverse-variance weights from a random-effects meta-analysis in order to estimate predictive ability in a new study.

Certain limitations of our proposed methods remain. Firstly, our approach does not combine estimates of predictive ability from nested case-control studies with those from cohort studies. Estimates from studies with a case-cohort design, however, may be more comparable with those from full cohorts (34). Secondly, persons with missing predictors are often excluded. Multiple imputation methods (43) applicable to multiple studies need further investigation.

As the scientific benefits of meta-analysis of individual participant data become increasingly recognized, there are a growing number of collaborative consortia being established. The methods presented in this paper provide practical solutions for assessment of the overall predictive ability of risk models, as well as the added value of novel predictors, in such collaborative enterprises.

ACKNOWLEDGMENTS

Author affiliations: Strangeways Research Laboratory, Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom (Lisa Pennells, Stephen Kaptoge, Simon G. Thompson, Angela M. Wood); and MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom (Ian R. White).

This work was supported the United Kingdom Medical Research Council (grant G0701619 and Unit Programme U105260558). The Emerging Risk Factors Collaboration Coordinating Centre was supported by the British Heart Foundation (grant RG/08/014), the Medical Research Council, the United Kingdom National Institute of Health Research Cambridge Biomedical Research Centre, a specific grant from the Bupa Foundation, and an unrestricted educational grant from GlaxoSmithKline. Various sources have supported recruitment, follow-up, and laboratory measurements in the cohorts contributing to the Emerging Risk Factors Collaboration. Investigators in several of these studies have contributed to a list of relevant funding sources (http://ceu.phpc.cam.ac.uk/research/erfc/studies/).

Members of the Emerging Risk Factors Collaboration are listed in the Appendix.

Conflict of interest: none declared.

REFERENCES

1. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med.* 1999;130(6):515–524.

2. Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat Med*. 2004;23(6):907–926.
3. Thompson S, Kaptoge S, White I, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *Int J Epidemiol*. 2010;39(5):1345–1359.
4. The Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. *Stat Med*. 2009;28(3):389–411.
5. Asia Pacific Cohort Studies Collaboration. Determinants of cardiovascular disease in the Asia Pacific region: protocol for a collaborative overview of cohort studies. *Cardiovasc Dis Prev*. 1999;2:281–289.
6. Beral V, Bull D, Doll R, et al. Breast cancer and abortion: collaborative reanalysis of data from 53 epidemiological studies, including 83 000 women with breast cancer from 16 countries. *Lancet*. 2004;363(9414):1007–1016.
7. Bingham S, Riboli E. Diet and cancer—the European Prospective Investigation into Cancer and Nutrition. *Nat Rev Cancer*. 2004;4(3):206–215.
8. Thompson D, Easton DF, Breast Cancer Linkage Consortium. Cancer incidence in BRCA1 mutation carriers. *J Natl Cancer Inst*. 2002;94(18):1358–1365.
9. Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol*. 2008;37(2):234–244.
10. Lewington S, Whitlock G, Clarke R, et al. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. *Lancet*. 2007;370(9602):1829–1839.
11. Smith-Warner SA, Spiegelman D, Ritz J, et al. Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. *Am J Epidemiol*. 2006;163(11):1053–1064.
12. Uitterlinden AG, Ralston SH, Brandi ML, et al. The association between common vitamin D receptor gene variations and osteoporosis: a participant-level meta-analysis. *Ann Intern Med*. 2006;145(4):255–264.
13. Matsushita K, Mahmoodi BK, Woodward M, et al. Comparison of risk prediction using the CKD-EPI equation and the MDRD study equation for estimated glomerular filtration rate. *JAMA*. 2012;307(18):1941–1951.
14. Danesh J, Erqou S, Walker M, et al. The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. *Eur J Epidemiol*. 2007;22(12):839–869.
15. Wormser D, Kaptoge S, Di AE, et al. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. The Emerging Risk Factors Collaboration. *Lancet*. 2011;377(9771):1085–1095.
16. The Emerging Risk Factors Collaboration. Lipid-related markers and cardiovascular disease prediction. *JAMA*. 2012;307(23):2499–2506.
17. Kaptoge S, Di Angelantonio E, Pennells L, et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. The Emerging Risk Factors Collaboration. *N Engl J Med*. 2012;367(14):1310–1320.
18. Cox DR. Regression models and life-tables [with discussion]. *J R Stat Soc Ser B*. 1972;34(2):187–220.
19. White IR. Multivariate random-effects meta-analysis. *Stata J*. 2009;9(1):40–56.
20. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Stat Med*. 2011;30(20):2481–2498.
21. Leandro G. *Meta-Analysis in Medical Research: The Handbook for the Understanding and Practice of Meta-Analysis*. Oxford, United Kingdom: Blackwell Publishing Ltd; 2005.
22. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–387.
23. Harrell FE Jr, Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543–2546.
24. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med*. 2004;23(5):723–748.
25. Gonen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005;92(4):965–970.
26. Uno H, Cai T, Pencina MJ, et al. On the *C* statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30(10):1105–1117.
27. Newson R. Confidence intervals for rank order statistics and their differences. *Stata J*. 2006;6(3):309–334.
28. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539–1558.
29. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011;342:d549.
30. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559–1573.
31. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am J Epidemiol*. 1997;145(1):72–80.
32. Pencina MJ, Larson MG, D'Agostino RB. Choice of time scale and its effect on significance of predictors in longitudinal studies. *Stat Med*. 2007;26(6):1343–1359.
33. Thiebaut AC, Benichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med*. 2004;23(24):3803–3820.
34. Ganna A, Reilly M, de FU, et al. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. *Am J Epidemiol*. 2012;175(7):715–724.
35. Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics*. 2008;64(1):1–9.
36. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157–172.
37. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30(1):11–21.
38. Royston P. Explained variation for survival models. *Stata J*. 2006;6(1):83–96.
39. Hosmer DW Jr, Lemeshow S. *Applied Logistic Regression*. New York, NY: John Wiley & Sons, Inc; 1989.
40. Parzen M, Lipsitz SR. A global goodness-of-fit statistic for Cox regression models. *Biometrics*. 1999;55(2):580–584.
41. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*.

New York, NY: Springer Science+Business Media, LLC; 2009.
42. Phillips AN, Thompson SG, Pocock SJ. Prognostic scores for detecting a high risk group: estimating the sensitivity when applied to new data. *Stat Med*. 1990;9(10):1189–1198.
43. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.

APPENDIX

Emerging Risk Factors Collaboration investigators/contributors—Air Force/Texas Coronary Atherosclerosis Prevention Study: Robert W. Tipping; Atherosclerosis Risk in Communities Study: Aaron R. Folsom, David J. Couper, Christie M. Ballantyne, Josef Coresh; British Regional Heart Study: S. Goya Wannamethee, Richard W. Morris; Bruneck Study: Stefan Kiechl, Johann Willeit, Peter Willeit, Georg Schett; British Women's Heart and Health Study: Shah Ebrahim, Debbie A. Lawlor; Caerphilly Prospective Study: John W. Yarnell, John Gallacher; Cardiovascular Health Study: Mary Cushman, Bruce M. Psaty, Russ Tracy (see http://www.chs-nhlbi.org for acknowledgements); Copenhagen City Heart Study: Anne Tybjærg-Hansen, Ruth Frikke-Schmidt, Marianne Benn, Børge G. Nordestgaard; Edinburgh Artery Study: Jackie F. Price, Amanda J. Lee, Stela McLachlan; European Prospective Investigation of Cancer–Norfolk Study: Kay-Tee Khaw, Nicholas J. Wareham; Epidemiologische Studie zu Chancen der Verhütung, Früherkennung und Therapie chronischer Erkrankungen in der älteren Bevölkerung: Hermann Brenner, Ben Schöttker, Heiko Müller, Dietrich Rothenbacher; First Myocardial Infarction in Northern Sweden: Jan-Håkan Jansson, Patrik Wennberg; Finrisk Cohort 1992, Finrisk Cohort 1997: Veikko Salomaa, Kennet Harald, Pekka Jousilahti, Erkki Vartiainen; Fletcher Challenge Blood Study: Mark Woodward; Framingham Offspring Study: Ralph B. D'Agostino, Sr., Philip A. Wolf, Ramachandran S. Vasan, Emelia J. Benjamin; Research Centre for Prevention and Health: Else-Marie Bladbjerg, Torben Jørgensen, Lars Møller, Jørgen Jespersen; Hisayama Study: Yutaka Kiyohara, Hisatomi Arima, Yasufumi Doi, Toshiharu Ninomiya; Hoorn Study: Jacqueline M. Dekker, Giel Nijpels, Coen D. A. Stehouwer; Kuopio Ischaemic Heart Disease Study: Jussi Kauhanen, Jukka T. Salonen; Lower Extremity Arterial Disease Event Reduction Trial: Tom W. Meade, Jackie A. Cooper; Multi-Ethnic Study of Atherosclerosis: Mary Cushman, Aaron R. Folsom, Bruce M. Psaty, Steven Shea (see http://www.mesa-nhlbi.org for acknowledgements); MONICA/KORA Augsburg Surveys S1, S2, and S3: Angela Döring, Wolfgang Koenig, Christa Meisinger; Multiple Risk Factor Intervention Trial 1: Lewis H. Kuller, Greg Grandits; Third National Health and Nutrition Examination Survey: Richard F. Gillum, Michael Mussolino; Nurses' Health Study: Eric B. Rimm, Sue E. Hankinson, JoAnn E. Manson, Jennifer K. Pai; Nova Scotia Health Survey: Susan Kirkland, Jonathan A. Shaffer, Daichi Shimbo; Prevention of Renal and Vascular End Stage Disease Study: Stephan J. L. Bakker, Ron T. Gansevoort, Hans L. Hillege; Prospective Epidemiological Study of Myocardial Infarction: Philippe Amouyel, Dominique Arveiler, Alun Evans, Jean Ferrières; Prospective Study of Pravastatin in the Elderly at Risk: Naveed Sattar, Rudi G. Westendorp, Brendan M. Buckley; Quebec Cardiovascular Study: Bernard Cantin, Benoît Lamarche, Jean-Pierre Després, Gilles R. Dagenais; Rancho Bernardo Study: Elizabeth Barrett-Connor, Deborah L. Wingard, Richele Bettencourt; Reykjavik Study: Vilmundur Gudnason, Thor Aspelund, Gunnar Sigurdsson, Bolli Thorsson; Rotterdam Study: Maryam Kavousi, Jacqueline C. Witteman, Albert Hofman, Oscar H. Franco; Strong Heart Study: Barbara V. Howard, Ying Zhang, Lyle Best, Jason G. Umans; Turkish Adult Risk Factor Study: Altan Onat; Uppsala Longitudinal Study of Adult Men: Johan Sundström, Liisa Byberg, Karl Michaëlsson; US Physicians Health Study: J. Michael Gaziano, Meir Stampfer, Paul M. Ridker; US Physicians Health Study 2: J. Michael Gaziano, Paul M. Ridker; Whitehall I Study: Michael Marmot, Robert Clarke, Rory Collins, Astrid Fletcher; Whitehall II Study: Eric Brunner, Martin Shipley, Mika Kivimäki; Women's Health Study: Paul M. Ridker, Julie Buring, Nancy Cook; West of Scotland Coronary Prevention Study: Ian Ford, James Shepherd, Stuart M. Cobbe, Michele Robertson. *Data management team*—Matthew Walker, Sarah Watson. *Coordinating center*—Myriam Alexander, Adam S. Butterworth, Emanuele Di Angelantonio, Pei Gao, Philip Haycock, Stephen Kaptoge, Lisa Pennells, Simon G. Thompson, Matthew Walker, Sarah Watson, Ian R. White, Angela M. Wood, David Wormser, John Danesh (Principal Investigator).