ONLINE FIRST

Predicting Death

An Empirical Evaluation of Predictive Tools for Mortality

George C. M. Siontis, MD; Ioanna Tzoulaki, PhD; John P. A. Ioannidis, MD, DSc

Background: The ability to predict death is crucial in medicine, and many relevant prognostic tools have been developed for application in diverse settings. We aimed to evaluate the discriminating performance of predictive tools for death and the variability in this performance across different clinical conditions and studies.

Methods: We used Medline to identify studies published in 2009 that assessed the accuracy (based on the area under the receiver operating characteristic curve [AUC]) of validated tools for predicting all-cause mortality. For tools where accuracy was reported in 4 or more assessments, we calculated summary accuracy measures. Characteristics of studies of the predictive tools were evaluated to determine if they were associated with the reported accuracy of the tool.

Results: A total of 94 eligible studies provided data on 240 assessments of 118 predictive tools. The AUC ranged from 0.43 to 0.98 (median [interquartile range], 0.77 [0.71-0.83]), with only 23 of the assessments reporting excellent discrimination (10%) (AUC, >0.90). For 10 tools, accuracy was reported in 4 or more assessments; only 1 tool had a summary AUC exceeding 0.80. Established tools showed large heterogeneity in their performance across different cohorts (I^2 range, 68%-95%). Reported AUC was higher for tools published in journals with lower impact factor ($P=.01$), with larger sample size ($P=.01$), and for those that aimed to predict mortality among the highest-risk patients ($P=.002$) and among children ($P<.001$).

Conclusions: Most tools designed to predict mortality have only modest accuracy, and there is large variability across various diseases and populations. Most proposed tools do not have documented clinical utility.

Arch Intern Med.
Published online July 25, 2011.
doi:10.1001/archinternmed.2011.334

Author Affiliations:
Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece (Drs Siontis, Tzoulaki, and Ioannidis); Department of Epidemiology and Biostatistics, Imperial College of Medicine, London, England (Drs Tzoulaki and Ioannidis); the Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts (Dr Ioannidis); the Department of Epidemiology, Harvard School of Public Health, Boston (Dr Ioannidis); and the Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, California (Dr Ioannidis).

THE ABILITY TO PREDICT death accurately is crucial for conveying information to patients about their future; for making sound medical decisions for management, treatment, and prevention; and for having realistic expectations. Evidence suggests that physicians perform poorly in predicting when patients will die.[1,2] However, numerous models have been developed to predict mortality in diverse settings.[3-5]

Herein we aim to empirically evaluate the ability of available predictive tools (multivariate or single variables) to predict the risk of death accurately for diverse conditions and populations. We assess how accurately and consistently these tools perform to help understand their potential clinical utility.

METHODS

SEARCH STRATEGY

To evaluate recently published studies that assessed the accuracy (discrimination) of tools to predict mortality, we searched Medline for studies published in 2009 by using the Clinical Queries tool. For more details on our search strategy and data extraction, see the eAppendix (www.archinternmed.com).

STUDY SELECTION

We included studies of any design published in 2009 that assessed the accuracy of tools to predict mortality (either single predictors or multivariable models); included assessment of accuracy based on the area under the receiver operating characteristic curve (AUC) (aka, C statistic or C index); and focused on all-cause death as the primary outcome. The AUC[6-9] is the most commonly used metric for assessing the accuracy of predictive tools.[10] The AUCs can be compared across different tools, while relative risk metrics depend on the unit to which they are expressed and cannot directly compare predictive tools expressed for different units of measurement.[11]

We excluded studies that only had data on the development of a new predictive tool or validated the predictive tool in the same cohort where it was developed because new, nonvalidated predictive tools are likely to have inflated estimates of accuracy.[12-14] We also excluded articles that did not provide primary data (eg, reviews) and studies where death was part

of a composite outcome or was determined as cause-specific (rather than all-cause) mortality.

When there were several eligible predictive tools and/or they assessed the ability to predict death at different lengths of follow-up in the same cohort, each proposed predictive tool and each time of follow-up assessment was included separately. For example, one study examined 2 predictive tools (Multidimensional Prognostic Index [MPI] and Pneumonia Severity Index [PSI]) for a total of 6 assessments at 3 different follow-up periods (1, 6, and 12 months) (See reference S47 in eReferences).

DATA EXTRACTION

The full text of the eligible studies and any supplementary materials were scrutinized to extract information on study design, characteristics of the cohort (prevalence of specific diseases), characteristics of the predictive tool and data on calibration,[15] reclassification,[16-18] and accuracy. For each study, we recorded the journal impact factor per the Institute for Scientific Information.[19] Calibration examines whether the risk prediction is equally good for patients at different levels of risk or there is a lack of fit. Reclassification examines whether the predictive tool helps classify patients in different, more appropriate risk categories compared with what could be done without its knowledge or compared with some other model. Accuracy is assessed by the AUC.

ANALYSIS

The AUC was defined as mean (SD) or median (interquartile range [IQR]). An AUC of 1 indicates perfect discrimination, while an AUC of 0.5 indicates discrimination no better than chance. While there are no absolute thresholds, usually an AUC of greater than 0.80 is considered to show very good discrimination, and AUC greater than 0.90 suggests excellent discrimination.[9]

For predictive tools where there was more than 1 assessment available, we noted the range of AUC values. For predictive tools with at least 4 data sets where both the AUC and corresponding 95% confidence intervals (CIs) were available, we summarized the AUC estimates using random effect models, weighting the AUC of each data set by the inverse of the sum of the between and within-study variances.[20-22] We quantified the heterogeneity in AUC values by the I^2 metric and its 95% CI. The I^2 metric takes values between 0% and 100%, and it is independent of the number of data sets (50%-75% indicates moderate heterogeneity, while >75% indicates very large heterogeneity).[23]

We compared the AUC values among prespecified subgroups based on prevalence of disease and predictive tool characteristics using 1-way analysis of variance for categorical variables and the Spearman correlation coefficient for continuous variables. Analyses were performed with STATA software, version 10.0 (StataCorp LP, College Station, Texas).

RESULTS

ELIGIBLE STUDIES AND PREDICTIVE TOOLS

Overall 544 items were retrieved from Medline, of which 235 were reviewed in full text. Of those, 94 articles (eReferences) were deemed eligible (eFigure). The interrater agreement (between G.C.M.S. and I.T.) for the selection of the eligible studies had κ value of 0.86.

These 94 manuscripts presented data on 240 assessments (224 multivariate models and 16 single predictors) of the accuracy of 118 predictive tools. Characteristics of studies and predictive tool assessments are listed in eTable 1. Most of the studies were performed in the United States or Europe, had a prospective cohort design, and pertained to acute disease conditions. Cardiovascular, critical-illness, infectious, gastroenterology-related, and malignant diseases accounted for 83% of the cohorts, but many other diseases were also assessed (eTable 1, eTable 2, and eTable 3). The median (IQR) sample size for the assessments was 502 (185-2016); the median (IQR) number of deaths was 71 (32-157); the median (IQR) proportion of deaths was 14% (5%-29%); and the median (IQR) death rate was 13% (4%-44%) per month. Among the whole data set (94 studies), in only 1 study (S85 in eReferences) did the investigators review and abstract patient data blinded to patients' hospital course and clinical status (eTable 1). For 78 studies, the percentage of losses to follow-up was available (70 studies reported no losses, while for the rest loss was generally low (median [IQR] loss to follow-up, 3.5% [1.25%-10.25%]).

PREDICTIVE TOOLS

Overall, 110 different predictive models and 8 different predictors were examined in the 240 assessments. The most commonly evaluated models included the Acute Physiology And Chronic Health Evaluation (APACHE) II model (n=19) and the MELD score (Model for End-Stage Liver Disease) (n=17) (**Table 1**). The predictive models included a wide range of variables (eTable 2). The number of variables in the models ranged from 2 to 30, and the median (IQR) number was 6 (4-12). All of the identified single predictors were biomarkers (eTable 3).

CALIBRATION AND RECLASSIFICATION

Calibration of the examined predictive tools was examined in fewer than half of the included studies (n=45; 48%), mainly by using the Hosmer-Lemeshow statistic (n=35; 78%) and observed/predicted ratio (n=5; 11%). Results were available in 44 studies (105 predictive tool assessments), indicating lack of fit for 8 studies (17 predictive tools).

Only 1 study (S83 in eReferences) examined reclassification analysis by means of the net reclassification improvement and the integrated discrimination index. This study investigated the added predictive value of radiographic ascites over and above the MELD-Na score in patients with cirrhosis.

ACCURACY

The AUC values ranged from 0.43 to 0.98 (**Figure 1**), and the median (IQR) AUC value was 0.77 (0.71-0.83). A total of 95 of the AUC values were higher than 0.80 (very good discrimination) (40%), but only 23 were higher than 0.90 (excellent discrimination) (10%).

The AUC data for all predictive tools with 2 or more assessments are listed in Table 1. For each of these 34 tools, the range of AUC estimates was large, sometimes spanning the spectrum from inaccurate to excellent ac-

Table 1. AUC Values of Predictive Tools Examined in More Than 1 Assessment

Predictive Tool[a]	Assessments, No.	AUC	
		Median (IQR)	Range
AMIS model	2	0.86 (0.84-0.87)	0.84-0.87
APACHE II	19	0.77 (0.71-0.81)	0.69-0.94
BCLC score	2	0.85 (0.84-0.86)	0.84-0.86
BISAP score	2	0.82 (NA)	0.82-0.82
BNP	3	0.66 (0.63-0.69)	0.63-0.69
CLIP score	5	0.88 (0.64-0.88)	0.62-0.96
CRIB II	2	0.91 (0.90-0.92)	0.90-0.92
CTP score	11	0.73 (0.72-0.84)	0.61-0.88
CURB-65 score	5	0.78 (0.73-0.78)	0.64-0.82
CCI	3	0.67 (0.63-0.74)	0.63-0.74
EuroSCORE	6	0.74 (0.70-0.77)	0.70-0.80
ISS	2	0.63 (0.54-0.72)	0.54-0.72
Intermountain risk score	3	0.87 (0.84-0.87)	0.84-0.87
JIS	5	0.85 (0.64-0.87)	0.59-0.87
MELD score	17	0.81 (0.78-0.86)	0.77-0.89
MELD-Na score	4	0.81 (0.78-0.86)	0.77-0.89
MESO index	3	0.87 (0.69-0.88)	0.69-0.88
MPI	3	0.80 (0.79-0.83)	0.79-0.83
MPM II	2	0.73 (0.66-0.79)	0.66-0.79
NT-pro-BNP	6	0.74 (0.71-0.76)	0.67-0.77
Pediatric death prediction model	2	0.92 (0.91-0.94)	0.91-0.94
PSI	7	0.75 (0.69-0.81)	0.63-0.83
Procalcitonin	2	0.73 (0.65-0.81)	0.65-0.81
RIFLE classification	3	0.75 (0.70-0.91)	0.70-0.91
Ranson's criteria	2	0.89 (0.82-0.95)	0.82-0.95
SAPS II	8	0.77 (0.73-0.82)	0.51-0.85
SAPS III	3	0.74 (0.71-0.84)	0.71-0.84
SOFA score	9	0.84 (0.75-0.85)	0.71-0.93
Simple risk index	2	0.80 (0.78-0.82)	0.78-0.82
TIMI risk score	5	0.73 (0.72-0.75)	0.68-0.84
TIMI risk score + laboratory index	2	0.77 (0.76-0.78)	0.76-0.78
TNM	2	0.80 (NA)	0.80-0.80
TRISS	2	0.75 (0.64-0.85)	0.64-0.85
Tokyo score	2	0.87 (0.86-0.87)	0.86-0.87

Abbreviations: AMIS, Acute Myocardial Infarction in Switzerland; APACHE II, Acute Physiology And Chronic Health Evaluation II; AUC, area under the receiver operating characteristic curve; BCLC, Barcelona Clinic Liver Cancer; BISAP, Bedside Index for Severity in Acute Pancreatitis; BNP, B-type natriuretic peptide; CCI, Charlson Comorbidity Index; CLIP, Cancer of the Liver Italian Program; CRIB II, Clinical Risk Index for Babies; CTP, Child-Turcotte-Pugh; CURB-65, confusion–blood urea nitrogen–respiratory rate–blood pressure–age ≥65 years; EuroSCORE, European system for cardiac operative risk evaluation; IQR, interquartile range; ISS, Injury Severity Score; JIS, Japan Integrated Staging; MELD, Model for End-Stage Liver Disease; MESO, MELD to SNa ratio; MPI, Multidimensional Prognostic Index; MPM II, Mortality Probability Models; NT-pro-BNP, *N*-terminal-pro-B-type natriuretic peptide; NA, not applicable; PSI, Pneumonia Severity Index; RIFLE, Risk of renal failure, injury to the kidney, failure of kidney function, loss of kidney function, and end-stage renal disease; SAPS, Simplified Acute Physiology Score; SOFA, Sequential Organ Failure Assessment; TIMI, Thrombolysis In Myocardial Infarction; TNM, tumor-nodes-metastasis; TRISS, Trauma Revised Injury Severity Score.

[a]To obtain further information about the specific studies that contribute AUC estimates to each predictive tool listed in this table, please contact the authors or consult the eTables and eReferences.

curacy. The median AUC values suggested modest accuracy. For only 2 predictive tools (Clinical Risk Index for Babies [CRIB] II [S25 and S27 in eReferences] and Pediatric death prediction model [S92 in eReferences]),

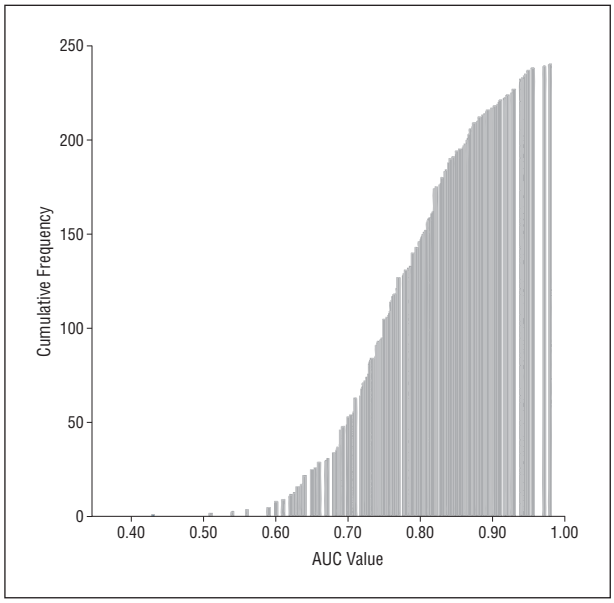

Figure 1. Cumulative frequency histogram of area under the receiver operating characteristic curve (AUC) values for mortality.

the median AUC value suggested excellent accuracy (AUC, 0.91 and 0.92, respectively), but this was based on only 2 assessments of each tool. Four or more assessments of the accuracy of a predictive tool were available for only 9 tools (APACHE, MELD, SOFA [Sequential Organ Failure Assessment], CTP [Child-Turcotte-Pugh], SAPS [Simplified Acute Physiology Score] II, PSI, CLIP [Cancer of the Liver Italian Program], CURB-65 [confusion–blood urea nitrogen–respiratory rate–blood pressure–age ≥65 years], JIS [Japan Integrated Staging]) and 1 biomarker (NT-pro-BNP [*N*-terminal-pro-B-type natriuretic peptide]). Using random effects meta-analysis, we found that the summary AUC estimates for these 10 tools ranged between 0.73 and 0.84 (**Figure 2**). For each of the 9 multivariable tools, there was marked heterogeneity of AUC values across diverse settings and studies (heterogeneity I^2 estimates in AUC ranged from 68% to 95%). The 95% CIs of the I^2 were also consistent with a large or very large heterogeneity. For NT-pro-BNP, the I^2 estimate was 25%. Meta-analyses retaining only the longest follow-up assessment when several follow-up assessments were available from the same study showed similar results (all changes in summary AUC estimates were <5% compared with the primary analysis including all data).

CORRELATES OF ACCURACY

As listed in **Table 2**, predictive tools published in journals of lower impact factor had higher reported AUC estimates than those published in journals of higher impact factor. Predictive tools were more accurate in predicting mortality when a smaller proportion of study participants died. The AUC values were also higher in pediatric than in adult populations. Finally, studies with larger sample size tended to have higher AUC values than smaller studies.

There was no evidence that study design (retrospective vs prospective), area of origin, disease status, clini-

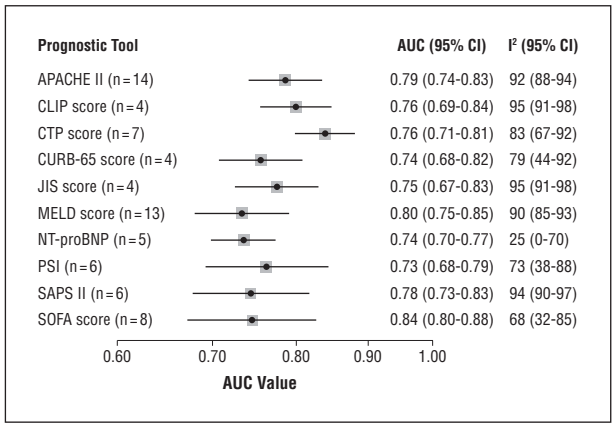

Figure 2. Area under the receiver operating characteristic curve (AUC) values for predictive tools that were examined in 4 or more assessments (n=number of assessments) with 95% confidence intervals (CIs). Summary results of AUC and 95% CIs are provided using random effects meta-analysis. APACHE II indicates Acute Physiology And Chronic Health Evaluation II; CLIP, Cancer of the Liver Italian Program; CTP, Child-Turcotte-Pugh; CURB-65, confusion–blood urea nitrogen–respiratory rate–blood pressure–age ≥65 years; JIS, Japan Integrated Staging; MELD, Model for End-Stage Liver Disease; NT-pro-BNP, N-terminal-pro-B-type natriuretic peptide; PSI, Pneumonia Severity Index; SAPS, Simplified Acute Physiology Score; SOFA, Sequential Organ Failure Assessment. To obtain further information about the specific studies that contribute AUC estimates to each predictive tool listed in this table, please contact the authors or consult the eTables and eReferences.

cal condition examined, death rate per month, loss to follow-up, or number of variables included in the predictive tool were associated with the AUC values (data not shown).

Table 2. Association Between AUC Values and Study Characteristics

Study Characteristic	All Predictive Tools		
	No.[a]	AUC, Mean (SD)	*P* Value[b]
Journal impact factor	222	NA	.021
≤2.13	46	0.78 (0.11)	
2.14-2.32	45	0.79 (0.07)	
2.33-3.15	45	0.78 (0.08)	
3.16-5.39	43	0.77 (0.07)	
>5.39	43	0.75 (0.10)	
Study population	240	NA	<.001
Pediatric	7	0.92 (0.02)	
Adult	225	0.77 (0.09)	
Both	8	0.78 (0.04)	
Sample size	240	NA	.01
≤147	48	0.76 (0.11)	
148-287	49	0.76 (0.11)	
288-810	48	0.76 (0.08)	
811-2558	48	0.80 (0.09)	
>2558	47	0.79 (0.08)	
Proportion of study participants who died	238	NA	.002
≤0.06	49	0.82 (0.08)	
0.07-0.13	47	0.76 (0.10)	
0.14-0.21	46	0.78 (0.10)	
0.22-0.33	50	0.78 (0.06)	
>0.33	46	0.73 (0.10)	

Abbreviations: AUC, area under the receiver operating characteristic curve; NA, not applicable.
[a] Number of the predictive tools related to the respective extracted variable.
[b] One-way analysis of variance for categorical variables (study population) and Spearman correlation coefficient for continuous variables (impact factor, sample size, and proportion of death).

COMMENT

Our systematic evaluation of a large number of seemingly well-validated predictive tools reported in the recent literature shows that these tools are not very accurate and that there is wide variation in their predictive accuracy for death. Most of the tools included in our analysis are not sufficiently accurate for wide use in clinical practice. Moreover, calibration was assessed in fewer than half of the tools, and of those tested, several showed lack of fit, meaning that prediction was not equally good for patients at different levels of risk. Studies published in journals with lower impact factor tended to show better AUC values, while tools performed better when they tried to predict death only for the highest-risk patients.

For a proposed predictive tool to be useful in clinical practice, there are several prerequisites. The tool must be validated in populations other than the one in which it was developed; it should be reproducible; and it should have good accuracy and calibration. Such a predictive tool can make accurate predictions in diverse settings across the range of both low- and high-risk patients. Few tools for predicting risk of death currently fit these criteria. Even tools that meet these criteria may not necessarily result in improvement in patient management and outcomes. This depends on whether effective, feasible interventions are available, the use of which is based on accurate knowledge of patient risk. However, reclassification, the ability to reclassify individuals into more appropriate risk categories where different actions/interventions might be indicated, is almost never assessed in the current literature of death prediction. Moreover, randomized trials on the use of predictive models, the ultimate proof of benefit, are few and difficult to conduct. Finally, clinicians are unlikely to use complex tools that require collection of extensive information, including data derived from expensive tests. It is possible that other predictive tools, based on far more limited clinical data, may perform equally well or better. In our empirical evaluation, models with more variables did not seem to perform clearly better than models with few variables.

Some characteristics of predictive tools were significantly associated with higher AUC estimates. For example, tools performed better when they tried to predict death only for the highest-risk patients. Excellent performance was seen in a small number of pediatric tools, while performance was substantially worse in predictive tools for adults. Larger studies tended to have slightly higher AUC estimates. These associations are exploratory and should be viewed with caution.

In our evaluation we focused on validated tools. However, even for some of the most widely applied predictive tools (such as APACHE II, MELD score, and SAPS II), we found great within-tool variability in accuracy across different studies and clinical settings. The observed variation of the accuracy for the same predictive

tool may be partly ascribed to the selective analysis and reporting of studies of predictive tools that may lead to exaggerated results of predictive discrimination in some studies. Efforts at standardization of reporting are important in this regard.[24,25] The inverse correlation between journal impact factor and reported AUC that we observed may represent lower methodologic quality with spuriously high reported predictive performance in some articles published in journals with low impact factor.[26] Moreover, studies often test predictive tools in populations that are very different than the one the model was developed for and for a wide range of outcomes. This may further contribute to the variability seen in their discriminatory performance.

Some limitations should be mentioned. Our empirical assessment was restricted to studies published during a single year. An effort to appraise the entire predictive literature would be a task requiring extensive international effort by hundreds of researchers, much as the Cochrane Collaboration has done for clinical trials. Moreover, we included only studies dealing with prediction of all-cause death, and we did not evaluate the accuracy of tools designed to predict other outcomes. However, death from any cause is a common outcome with great clinical impact, and it is possible to standardize unambiguously. Finally, we considered only predictive studies that assessed accuracy using the AUC. However, AUC is not the only metric to assess predictive ability,[27] and like any single metric, it can have limitations.[16,28-30] For example, the AUC does not provide information on the actual predicted probabilities, and it does not convey the exact risk distribution in the respective study population. Also, improvements in AUC are more difficult in the high-range values than when AUC is closer to 0.50.[6] Nevertheless, AUC is a very useful metric[16,30] and is the most widely used standardized metric in the predictive literature.

Given the very wide variability in the AUC, even for the same predictive tool, we believe that systematic efforts are needed to organize and synthesize the predictive literature, such as those proposed by the Cochrane Prognosis Methods Group. Such efforts are needed to enhance the evidence derived from predictive research and to establish standard methods for developing, evaluating, reporting,[31,32] and eventually adopting new predictive tools in clinical practice. Clinicians should be cautious about adopting new, initially promising predictive tools, especially complex ones based on expensive measurements that have not been extensively validated and shown to be consistently useful in practice.

Accepted for Publication: May 9, 2011.
Published Online: July 25, 2011. doi:10.1001/archinternmed.2011.334
Correspondence: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, Stanford University School of Medicine, 251 Campus Dr, MSOB X306, Stanford, CA 94305 (jioannid@stanford.edu).
Author Contributions: Dr Ioannidis had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. *Study concept and design:* Siontis, Tzoulaki, and Ioannidis. *Acquisition of data:* Siontis, Tzoulaki, and Ioannidis. *Analysis and interpretation of data:* Siontis, Tzoulaki, and Ioannidis. *Drafting of the manuscript:* Siontis, Tzoulaki, and Ioannidis. *Critical revision of the manuscript for important intellectual content:* Siontis, Tzoulaki, and Ioannidis. *Statistical analysis:* Siontis, Tzoulaki, and Ioannidis. *Administrative, technical, and material support:* Siontis, Tzoulaki, and Ioannidis. *Study supervision:* Ioannidis.
Financial Disclosure: None reported.

REFERENCES

1. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ*. 2000;320(7233):469-472.
2. Glare P, Virik K, Jones M, et al. A systematic review of physicians' survival predictions in terminally ill cancer patients. *BMJ*. 2003;327(7408):195-198.
3. Timsit JF, Fosse JP, Troché G, et al; OUTCOMEREA Study Group, France. Calibration and discrimination by daily Logistic Organ Dysfunction scoring comparatively with daily Sequential Organ Failure Assessment scoring for predicting hospital mortality in critically ill patients. *Crit Care Med*. 2002;30(9):2003-2013.
4. Krüger S, Ewig S, Giersdorf S, Hartmann O, Suttorp N, Welte T; German Competence Network for the Study of Community Acquired Pneumonia (CAPNETZ) Study Group. Cardiovascular and inflammatory biomarkers to predict short- and long-term survival in community-acquired pneumonia: results from the German Competence Network, CAPNETZ. *Am J Respir Crit Care Med*. 2010;182(11):1426-1434.
5. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*. 1997;336(4):243-250.
6. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*. 2008;54(1):17-23.
7. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;115(5):654-657.
8. Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med*. 2006;355(25):2631-2639.
9. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
10. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA*. 2009;302(21):2345-2352.
11. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882-890.
12. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453-473.
13. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54(8):774-781.
14. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? *Semin Urol Oncol*. 2002;20(2):96-107.
15. Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Stat Med*. 2002;21(18):2723-2738.
16. 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):157-172.
17. 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.
18. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med*. 2008;149(10):751-760.
19. Institute for Scientific Information. ISI web of knowledge. http://isiknowledge.com. Accessed April 15, 2011 [subscription required].
20. Kester AD, Buntinx F. Meta-analysis of ROC curves. *Med Decis Making*. 2000;20(4):430-439.
21. Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. *Stat Med*. 2009;28(3):389-411.
22. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127(9):820-826.
23. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558.

24. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst*. 2005;97(16):1180-1184.
25. Sigounas DE, Tatsioni A, Christodoulou DK, Tsianos EV, Ioannidis JP. New prognostic markers for outcome of acute pancreatitis: overview of reporting in 184 studies. *Pancreas*. 2011;40(4):522-532.
26. Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. *JAMA*. 2002; 287(21):2805-2808.
27. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
28. Diamond GA. What price perfection? calibration and discrimination of clinical prediction models. *J Clin Epidemiol*. 1992;45(1):85-89.
29. Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol*. 2010;172(3):353-361.
30. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935.
31. Riley RD, Sauerbrei W, Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. *Br J Cancer*. 2009; 100(8):1219-1229.
32. Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ; GRIPS Group. Strengthening the reporting of Genetic RIsk Prediction Studies: the GRIPS Statement. *PLoS Med*. 2011;8(3):e1000420.