Factors Associated With 5-Year Risk of Hip Fracture in Postmenopausal Women

John Robbins, MD
Aaron K. Aragaki, MS
Charles Kooperberg, PhD
Nelson Watts, MD
Jean Wactawski-Wende, MD
Rebecca D. Jackson, MD
Meryl S. LeBoff, MD
Cora E. Lewis, MD
Zhao Chen, PhD
Marcia L. Stefanick, PhD
Jane Cauley, DrPH

Context The 329 000 hip fractures that annually occur in the United States are associated with high morbidity, mortality, and cost. Identification of those at high risk is a step toward prevention.

Objective To develop an algorithm to predict the 5-year risk of hip fracture in postmenopausal women.

Design, Setting, and Participants A total of 93 676 women who participated in the observational component of the Women's Health Initiative (WHI), a multiethnic longitudinal study, were used to develop a predictive algorithm based on commonly available clinical features. Selected factors that predicted hip fracture were then validated by 68 132 women who participated in the clinical trial. The model was tested in a subset of 10 750 women who had undergone dual-energy x-ray absorptiometry (DXA) scans for bone mass density assessment.

Main Outcome Measure The prediction of centrally adjudicated hip fracture, measured by the area under the receiver operator characteristic (ROC) curves.

Results During a mean (SD) follow-up of 7.6 (1.7) years, 1132 hip fractures were identified among women participating in the observational study (annualized rate, 0.16%), whereas during a mean follow-up of 8.0 (1.7) years, 791 hip fractures occurred among women participating in the clinical trial (annualized rate, 0.14%). Eleven factors predicted hip fracture within 5 years: age, self-reported health, weight, height, race/ethnicity, self-reported physical activity, history of fracture after age 54 years, parental hip fracture, current smoking, current corticosteroid use, and treated diabetes. Receiver operating characteristic curves showed that the algorithm had an area under the curve of 80% (95% confidence interval [CI], 0.77%-0.82%) when tested in the cohort of different women who were in the clinical trial. A simplified point score was developed for the probability of hip fracture. Receiver operating characteristic curves comparing DXA-scan prediction based on a 10% subset of the cohort and the algorithm among those who participated the clinical trial were similar, with an area under the curve of 79% (95% CI, 73%-85%) vs 71% (95% CI, 66%-76%).

Conclusion This algorithm, based on 11 clinical factors, may be useful to predict the 5-year risk of hip fracture among postmenopausal women of various ethnic backgrounds. Further studies are needed to assess the clinical implication of the algorithm in general and specifically to identify treatment benefits.

JAMA. 2007;298(20):2389-2398 www.jama.com

THE ESTIMATED 329 000 HIP fractures that occur annually in the United States[1] are associated with high morbidity, mortality, and cost.[2] Prevention of hip fracture is a high priority for patients, physicians, and public health. Several studies and consensus opinions have investigated the risk factors for hip fractures.[3-8] The Study of Osteoporotic Fractures (SOF),[3] which included 7782 women over 5 years, set the benchmark for establishing risk of hip fracture to date. The number of women included in the Women's Health Initiative (WHI) is an order of magnitude larger than SOF, and WHI includes nearly 20% minority women.

Although dual-energy x-ray absorptiometry (DXA) scan can precisely predict risk of hip fractures, as it did for a small subset of women participating in WHI, by assessing bone mineral density (BMD), clinicians and patients would benefit from assessing risk by

See also Patient Page.

Author Affiliations: Department of Internal Medicine, University of California at Davis School of Medicine, Sacramento (Dr Robbins); Fred Hutchinson Cancer Research Center, Seattle, Washington (Mr Aragaki and Dr Kooperberg); University of Cincinnati College of Medicine, Cincinnati, Ohio (Dr Watts); Division of Endocrinology, Diabetes and Metabolism, University at Buffalo, Buffalo, New York (Dr Wactawski-Wende); Department of Internal Medicine and Physical Medicine, Ohio State University, Columbus (Dr Jackson); Department of Medicine, Harvard Medical School, Boston, Massachusetts (Dr LeBoff); Division of Preventive Medicine, University of Alabama at Birmingham, (Dr Lewis); Division of Epidemiology and Biostatistics, University of Arizona, Tucson (Dr Chen); Stanford Prevention Research Center, Stanford, California (Dr Stefanick); Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania (Dr Cauley).
Corresponding Author: John Robbins, MD, Department of Internal Medicine, University of California Davis School of Medicine, 4150 V St, Ste 2400, Sacramento, CA 95817 (jarobbins@ucdavis.edu).

other means. Most hip fractures occur in women who are not osteoporotic by BMD testing.[9] Furthermore, it has been suggested by Black et al[4] that an algorithm without BMD is nearly as predictive as one with BMD.

The purpose of our study was to evaluate clinical risk factors for hip fracture in a multiethnic cohort of more than 100 000 postmenopausal women. Our goal was to create and test a predictive model for hip fracture using the WHI cohorts. It is important to investigate the combined effects of risk factors for hip fracture. There is the potential problem of interpreting factors independently of each other. For example, prior studies that had associated the risk of hip fracture with specific ethnic groups may have identified a marker of risk not a cause because they failed to adjust for such factors as weight, smoking status, and other risks.[10] Only hip fracture risk was evaluated. By studying hip fractures, we were able to use data from medical records to clearly identify those with fractures. Had we included other fractures, such as spine fractures, we would have had to rely on self-report. Although spine factures result in significant morbidity and mortally, hip fractures are clearly more detrimental to a woman's health.

METHODS

Study Population

The WHI has multiple components that can be used to build and test a predictive algorithm by taking advantage of an overlapping multicomponent design. Thus, some women were in multiple intervention components of the study. The WHI recruited postmenopausal women aged 50 to 79 years from 40 clinical centers and assigned them to multiple clinical trial components and to an observational study. The dietary modification component included 48 835 eligible women who were randomly assigned to either a sustained low-fat eating pattern (40%) or to eat as they pleased (60%).[11] The hormone therapy clinical trial randomized 27 347 women to trials assessing estrogen plus progestin or estrogen alone compared with placebo; women who still had a uterus received 0.625 mg of conjugated equine estrogen and 2.5 mg of medroxyprogesterone acetate or placebo daily while women without a uterus received estrogen alone or placebo.[12] Approximately 1 year after randomization into 1 of the above components, 36 282 women in the hormone therapy and/or dietary modification trial were randomly assigned to receive 1 g of calcium plus 400 IU of cholecalciferol (vitamin D) or placebo daily.[13]

All of the participants, including those who agreed to being followed up after dropping out of the interventions, are used in this analysis. Mean follow-up of the participants varies by component. The study treatments in the 2 components of the hormone trial were stopped prematurely; however, women continued to be followed up for events until study close-out. Women in the estrogen plus progestin group discontinued intervention after a mean of 5.6 years. Women in the estrogen-only group were followed up while taking the study drugs for 6.8 years. The dietary modification intervention lasted a mean of 8.1 years. Follow-up in the calcium vitamin D trial was a mean of 7 years. Information on the study design, methods, and results of these trials has been previously reported.[3,14-19] The mean (SD) follow-up time for women in the clinical trial was 8.0 (1.7) years (median, 8.0 years; interquartile range, 7.4-9.0 years). The participants in WHI were generally healthier and had more education than the general US population of women in the same age range.[20]

Postmenopausal women who were screened for the clinical trial but were ineligible or unwilling to participate in randomization were asked to enroll in an observational study. Women were ineligible if they did not want to discontinue taking hormone therapy upon study entry, or had a history of breast cancer; they were ineligible for the dietary component if they already followed a low-fat diet or too frequently ate away from home; and they were ineligible for the calcium and vitamin D component if they had a history of kidney stones or were unwilling to limit vitamin D intake.[21] A total of 93 676 women who enrolled in the observational study, were evaluated for multiple risk factors and followed up for a mean (SD) of 7.6 (1.7) years (median, 7.9 years; interquartile range, 6.9-8.9 years). Similar questionnaires and methods were used to determine baseline characteristics for both the clinical trial and the observations study groups. A subset of WHI participants from 3 of 40 clinical sites underwent DXA scans.

Incidence of hip fracture was collected using a standardized medical update questionnaire completed by all participants. These were collected every 6 months for those in the clinical trial and annually for those in the observational study until the study closed between October 2004 and March 2005. Hip fractures were self-reported and then confirmed both locally and centrally by review of medical records including x-ray and surgical reports. Agreement rate between self-reported hip fracture and adjudicated results based on medical records review was good, 78%,[22] but not perfect, and substantiates the need for individual review of outcomes, not just self-report as has been used in a number of other studies. All of the protocols were approved by the appropriate institutional review boards and participants signed informed consents.

Variables

Most of the variables are self-explanatory. (For a complete list of procedures, see http://www.whiscience.org/about/about_collection.php) Height and weight were measured in the clinics with calibrated scales and stadiometers. Two blood pressure and pulse measurements were manually obtained by trained technicians after 5 minutes of rest at 30 seconds apart. Prevalent medical conditions and medications, eg, diabetes, corticosteroid use, were based on self-report. Physical activity was self-reported and measured

as metabolic equivalent tasks (METs), using values derived from the literature and standardized questionnaires, which were validated for reproducibility in this population.[23] Similar questions have been validated against exercise diaries.[24] A MET is the ratio of work metabolic rate to a standard resting metabolic rate of 4.184 kJ/kg per hour.[25] For example, activity intensity were coded as 7 METS for strenuous, 4 for moderate, and 3 for low. Mean walking speed was classed as 3 METS for a 2 to 3 mph, 4 for 3 to 4 mph, and 4.5 for 4 mph or faster. METs per week were calculated as MET-h/wk.

Risk for depression was obtained from the Centers for Epidemiologic Studies–Depression 6-item questionnaire.[26] (This is unrelated to medication or physician diagnosis.) Dietary data were collected via self-report using food frequency questionnaire.[27] Dietary quality was identified using the method described by Neuhouser et al.[28] In brief, dietary intakes of fat, saturated fat, cholesterol, fruit and vegetables, sodium, calcium, protein, and fiber were coded as a 0 if achieved dietary recommendation, 1 if achieved within 30% of dietary recommendation, and 2 for everything else. The 8 scores were then summed. Lower scores indicate a better diet. Race and ethnicity were self-identified by the participants.

Statistical Methods

A prediction model was developed from the WHI observational study dataset and validated by the WHI clinical trial dataset. The observational study population was much larger than the clinical trial and more heterogeneous, thus offered more power for the development of the algorithm.

Model Development

Potential risk factors were identified from the literature and fit 1 at a time in a Cox proportional hazards model, adjusting for age and race/ethnicity. Variables that achieved a modest level of statistical significance ($P<.25$), based on the score test, were included

Table 1. Baseline Characteristics by Hip Fracture During Follow-up in the Observational Study Cohort

Baseline Characteristic[a]	Incident Hip Fracture, No. (%)		*P* Value[b]
	No	Yes	
Age group at screening, y			
50-59	29 603 (32.0)	102 (9.0)	<.001
60-69	40 838 (44.1)	359 (31.7)	
70-79	22 103 (23.9)	671 (59.3)	
Race/ethnicity			
White	76 949 (83.1)	1064 (94.0)	<.001
Black	7612 (8.2)	27 (2.4)	
Hispanic	3612 (3.9)	11 (1.0)	
American Indian	417 (0.5)	5 (0.4)	
Asian/Pacific Islander	2660 (2.9)	11 (1.0)	
Unknown	1294 (1.4)	14 (1.2)	
Marital status			
Never married	4322 (4.7)	68 (6.0)	.04
Divorced/separated	14 593 (15.8)	134 (11.9)	
Widowed	15 964 (17.3)	326 (28.8)	
Presently married/living as married	57 203 (62.1)	602 (53.3)	
Has medical insurance	89 011 (97.2)	1110 (98.8)	.17
Physical activity (METs/wk)			
0, Inactive	12 456 (13.6)	181 (16.3)	<.001
<5	17 522 (19.1)	241 (21.7)	
5-12	21 559 (23.6)	292 (26.3)	
≥12	39 983 (43.7)	395 (35.6)	
Smoking status			
Never smoked	46 458 (50.9)	565 (50.6)	<.001
Past smoker	39 058 (42.8)	456 (40.8)	
Current smoker	5695 (6.2)	96 (8.6)	
Parent broke hip after age 40	12 403 (13.4)	240 (21.2)	<.001
Fracture on or after age 55 y			
No	60 728 (71.2)	655 (65.6)	<.001
Yes	12 228 (14.3)	313 (31.4)	
Not available	12 356 (14.5)	30 (3.0)	
Alcohol consumption, drinks/d			
Nondrinker	38 707 (41.9)	535 (47.3)	<.001
≤1	42 111 (45.6)	459 (40.6)	
>1	11 573 (12.5)	136 (12.0)	
Medication			
Supplemental calcium	55 264 (59.7)	670 (59.2)	.07
Antianxiety or antidepressant	9389 (10.1)	127 (11.2)	.04
Bisphosphonate	1989 (2.1)	45 (4.0)	.008
Oral daily corticosteroid	1162 (1.3)	41 (3.6)	<.001
Thyroid hormone	13 349 (14.4)	213 (18.8)	.08
Hormone therapy			
Never used	37 466 (40.5)	559 (49.4)	.005
Past user	13 721 (14.8)	199 (17.6)	
Current user	41 273 (44.6)	373 (33.0)	
Prior bilateral oophorectomy	18 699 (20.7)	192 (17.6)	.003
Age at menarche, y			
<12	20 328 (22.1)	197 (17.6)	.04
12-13	50 780 (55.1)	610 (54.5)	
≥14	21 041 (22.8)	313 (27.9)	
No. of term pregnancies			
Never pregnant or never had term pregnancy	11 775 (12.8)	163 (14.6)	.06
1-2	32 529 (35.4)	389 (34.9)	
>3	47 619 (51.8)	563 (50.5)	

(continued)

in the pool of variables used to select a final prediction model. Ten-fold cross-validation was used to determine the optimal number of predictors that minimizes an estimate of prediction error.[27,29] Specifically, we divided the training data into 10 parts. Nine-tenths of the data was used to select the best model with k predictors by fitting a hazard regression model, which uses stepwise addition and deletion and considers interactions and nonparametric (spline) terms. For each model, we then evaluated the prediction log-likelihood on the remaining one-tenth of the data that was not used to select the model. For each k, we added these predicted log likelihoods to obtain a prediction score. The value of k that minimizes the cross-validated prediction score is taken to be the optimal number of predictors. A hazard regression model with K* predictors was then selected from the entire WHI observational study data.

The probability of a hip fracture within 5 years was then calculated using a multivariate logistic regression model fit on the WHI observational study dataset, using the K* variables selected above. The Hosmer-Lemeshow statistic was used to ascertain lack-of-fit (calibration) of this model. Participants with missing data in their predictor variables, and 5.5% (n=5161) of the participants who did not have a hip fracture within 5 years or did not have 5 years of follow-up were excluded from the logistic regression model.

Table 1. Baseline Characteristics by Hip Fracture During Follow-up in the Observational Study Cohort (cont)

Baseline Characteristic[a]	Incident Hip Fracture, No. (%)		*P* Value[b]
	No	Yes	
>10 lb intentional weight loss in last 20 y	49 475 (53.9)	487 (43.4)	.003
Depressive symptom[c]			
0	23 679 (26.1)	273 (24.8)	.04
1-2	33 516 (36.9)	387 (35.1)	
3-4	19 038 (21.0)	259 (23.5)	
>5	14 580 (16.1)	(16.7)	
Baseline general			
Excellent	16 437 (17.9)	139 (12.4)	<.001
Very good	37 303 (40.6)	382 (34.1)	
Good	29 255 (31.8)	414 (37.0)	
Fair	8036 (8.7)	174 (15.5)	
Poor	872 (0.9)	10 (0.9)	
Treated diabetes	38 423 (4.1)	79 (7.0)	<.001
Diet quality index, quartile[d]			
1st	14 387 (16.2)	170 (15.8)	.17
2nd	23 284 (26.2)	285 (26.5)	
3rd	28 818 (32.4)	357 (33.2)	
4th	22 350 (25.2)	263 (24.5)	
	No. Mean (SD)	No. Mean (SD)	
Height, cm	91797 161.7 (6.8)	1123 161.8 (7.1)	<.001
Weight, kg	92077 71.7 (16.9)	1127 67.7 (15.5)	<.001
Dietary calcium, mg	88 839 778.8 (435.3)	1075 765.9 (445.4)	.06
Dietary vitamin D, μg	88 839 5.0 (3.2)	1075 5.0 (3.2)	.10
Change in height from age 18, %	89 612 −1.0 (3.3)	1097 −1.9 (3.6)	<.001
Change in weight from age 35, %	90 951 19.8 (22.3)	1120 12.6 (20.8)	<.001

[a] For brevity, baseline characteristics that did not have a modest marginal association hip fracture ($P > .25$), after adjusting for age and ethnicity, are not shown. These include: use of supplements containing cholecalciferol (vitamin D), multivitamins, thiazides and thiazidelike diuretic, hypnotic medication, benzodiazepines, antiestrogens, oral contraceptive use, age at menopause, resting pulse, education, cups of regular coffee, calcitonin use, age at first birth, and currently following lactose-free diet.

[b] *P* value corresponds to the marginal association of baseline characteristic with hip fracture. *P* value is from a Cox proportional hazards model adjusting for age and ethnicity. *P* values for age and ethnicity correspond to unadjusted marginal associations.

[c] Sum of Center for Epidemiologic Studies–Depression score. A higher score indicates greater depression.

[d] Dietary intakes of fat, saturated fat, cholesterol, fruit and vegetables, sodium, calcium, protein, and fiber were coded as a 0 if achieved dietary recommendation, 1 if achieved within 30% of dietary recommendation, and 2 otherwise. The 8 scores are then summed. Lower scores indicate a better diet.

Model Validation

To avoid an overly optimistic evaluation of model validity, we use the WHI clinical trial participants as our validation dataset. The women in the clinical trial were different in a multiple ways from the women in the observational study. The women in the clinical trial had volunteered to participate, were taking trial-required medications, and were following diet plans. These differences work to improve the usefulness of the validation; it is important that the algorithm work for women with different characteristics. The probability of a hip fracture within 5 years for the validation data was based on the multivariate logistic regression coefficients calculated exclusively on the WHI observational study data. Receiver operator characteristic (ROC) curves and the corresponding area under the curve (AUC) were used to evaluate how the prediction model preformed on the test data. The AUC was also calculated independently for the factors in the final model to demonstrate the additional value gained from the addition of each factor to the model. ROC curves plot the true-positive rate (sensitivity) vs the false-positive rate (1-specificity) at a continuum of thresholds; a participant is classified as having a hip fracture if her estimated probability of fracture exceeds a particular threshold. The ROC curve is a graphical representation of test characteristics, with sensitivity on the y-axis and 1-specificity on the x-axis, over all possible cut points for defining a positive and a negative test result. For our study, a positive result—predicting that an individual would have a hip fracture—occurs when the probability of fracture lies above a cut point.[30]

Because of the limited number of hip fractures in the DXA subset of women, a 10-fold cross-validation technique was used to compute the ROC curves and

AUC. The 95% confidence intervals (CIs) were obtained by bootstrapping.

Cox proportional hazards models, logistic regression models, and their corresponding statistics were computed using SAS version 9.1 (SAS Institute Inc, Cary, North Carolina). The hazard regression model fits, stepwise selection, cross-validation, and ROC/AUC were computed using R version 2.1 and R libraries polspline and ROCR (R Development Core Team, http://www.R-project.org).[31-33] $P<.05$ was considered statistically significant.

RESULTS

Over a mean (SD) follow up of 7.6 (1.7) years, women in the observational study experienced 1132 hip fractures, an annual rate of 0.16%, whereas during a mean follow-up of 8.0 (1.7) years 791 women in the clinical trial experienced hip fractures at an annual rate of 0.14%. The 10 750 women with BMD measurements were followed up for 5 years or until they fractured their hip. Eighty hip fractures occurred in the combined groups over a mean (SD) of 8.7 (1.2) years of follow-up. The variables considered for inclusion in the model are shown in TABLE 1. Variables that did not meet the nominal threshold ($P<.25$) for consideration were education; cups of regular coffee; age at menopause; age at first birth; maintaining a lactose-free diet; pulse pressure; intentional weight loss (≥4.5 kg [≥10 lbs]); and use of vitamin D supplements, multivitamins thiazides and thiazidelike diuretics, antihypnotics, benzodiazepines, antiestrogens, calcitonins, and oral contraceptives. The independent frequency or mean after adjustment for age and race/ethnicity in those with and without hip fracture and significance are included.

Development of Algorithm

Cross-validation and stepwise selection of hazard regression models identified 12 variables from Table 1 that were independently predictive of hip fracture. These variables were age, self-reported health, height, change in height since the age of 18 years, change in weight since the age of 35 years, history of fracture after the age of 55 years, race/ethnicity, physical activity, smoking, history of parental fracture after the age of 40 years, diabetes treated with medications, and corticosteroid use. We did not find any pairwise interactions or nonlinear terms that were predictive of hip fracture. The variables change in height since the age of 18 years and change in weight since the age of 35 years were not available for the WHI clinical trial test set that we had planned to use. We therefore chose to use weight as a surrogate for change in weight, this less-than-perfect substitution, errs on the conservative side (TABLE 2).

The participants who were excluded from the logistic regression model (who did not have a hip fracture within 5 years and who did not have 5 years of follow-up) tended to be minorities (28% vs 16%) and older age (66 vs 63 years). More than half of these women died before 5 years of follow-up (n=2768). The Hosmer-Lemeshow statistic indicated no sign of lack of fit ($P=.20$).

An interactive model is available at http://hipcalculator.fhcrc.org.

As a second step, we approximated the additive logistic regression model by multiplying the coefficients by an arbitrary constant (4, selected to yield approximately integer-valued additive factors) and rounded to the nearest integer. This yielded a simple additive score. The 5-year risk of hip fracture can be calculated by totaling the point score. A point total of 9 yields a probability of fracture of 0.1%, a point total of 18 yields a probability of fracture of 1%, and a point total of 24 yields a probability of fracture of 5%.

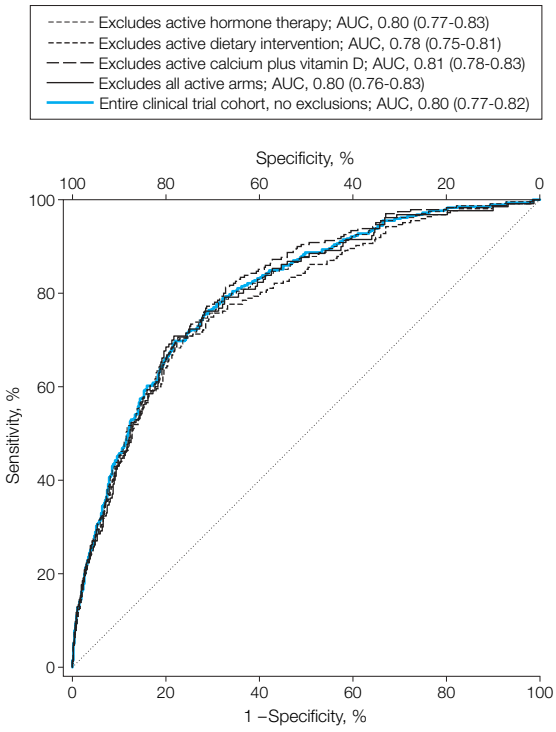
Validation

We tested the ability of the model to identify the 5-year probability of hip

Table 2. Multivariate Logistic Regression Model: Risk Factors for Hip Fracture in the Observational Study

Risk Factors	Odds Ratio (95% Confidence Interval)	*P* Value	Point Score
Age per each year	1.13 (1.11-1.15)	<.001	½ per year >50
Self-reported health			
Fair or poor vs excellent	2.38 (1.66-3.40)	<.001	3
Good vs excellent	1.22 (0.90-1.66)		1
Very good vs excellent	1.11 (0.83-1.49)		0
Height per each inch	1.11 (1.07-1.16)	<.001	½ per inch >64
Weight per each pound	0.99 (0.98-0.99)	<.001	1 per 25 lb <200
Fracture on or after age 55 y			
Not applicable vs no	1.01 (0.51-2.02)	<.001	0
Yes vs no	1.72 (1.41-2.10)		2
Race/ethnicity			White, 3
Unknown vs white	1.00 (0.47-2.14)	<.001	
Asian/Pacific Islander vs white	0.26 (0.10-0.70)		
American Indian vs white	1.60 (0.50-5.10)		
Hispanic vs white	0.32 (0.12-0.86)		
Black vs white	0.41 (0.24-0.70)		
Physical activity, METs			1
5-12 vs ≤12	1.32 (1.04-1.67)	.004	
<5 vs ≤12	1.26 (0.97-1.64)		
Inactive 0 vs ≤12	1.64 (1.24-2.17)		
Smoking status			
Current vs never	2.33 (1.71-3.18)	<.001	3
Past vs never	0.96 (0.79-1.17)		0
Parent broke hip, yes vs no	1.50 (1.20-1.87)	<.001	1
Corticosteroid use, yes vs no	1.94 (1.16-3.25)	.01	3
Use of hypoglycemic agent, yes vs no	1.74 (1.17-2.60)	.006	2

Figure 1. Women's Health Initiative Clinical Trial Test Set Receiver Operating Characteristic Curve



AUC indicates area under the curve. Blue curves in Figure 1 and Figure 2 are the same and are derived from the entire clinical trial cohort.

Table 3. Contributions of Individual Predictors

Variable	AUC%[a]
General health	56
Height	56
Weight	57
Fracture after age 55 y	56
Race/ethnicity	54
Physical activity	53
Currently smoking	53
Parent broke hip	51
Corticosteroid use	50
Diabetes	51
All predictors except age	67
Age	76[b]
Age plus all predictors	80[c]

Abbreviation: AUC, area under the curve.
[a] For predictor variables, other than age, weighted AUCs $\frac{\sum_i w_i AUC_i}{\sum_i w_i}$ are calculated where w_i is the number of hip fractures for ith age group and i goes from 50 to 79. The AUC_i are the AUCs for the ith age group. These are based on logistic regression model that contain the predictor of interest and age (categorical); trained on the observational study and tested on the clinical trial.
[b] Based on a logistic regression model containing age as a single variable; trained on the observational study and tested on the clinical trial.
[c] Based on our full logistic regression model; trained on the observational study and tested on the clinical trial.

fracture in women included in the hormone treatment, dietary, and calcium and vitamin D components of the WHI clinical trial. It should be noted that the women in the observational study cohort had different characteristics than those in clinical trial cohorts. Participants in the clinical trial tended to be younger (mean, 62.7 years), taller (161.1 cm [63.42 in]), heavier (76.1 kg [169.1 lb]), less likely to be white (81.5% were white), with a lower proportion of the clinical trial reporting fair to poor health (8.3%), history of fracture after age 55 years (13.1%), either parent breaking a hip (11.8%), and corticosteroid use (0.1%). A higher proportion of the clinical trial participants reported being physically inactive (19.2%), currently smoking (7.9%), and taking treatment for diabetes (4.8%). These differences between the clinical trial and observational study participants were all statistically significant ($P<.001$).

Using adjudicated hip fractures for women in the clinical trial, ROC curves were developed to test how well the algorithm that was developed from the observational study cohort performed in validation populations. The AUC was tested against the WHI clinical trial. We examined various groups participating in the clinical trial and found similar results in cases in which the AUC ranged from 78% to 81%. The AUC was 80% for all WHI clinical trial participants, all WHI participants receiving placebos, and those who received no active HT intervention (FIGURE 1).

Although there are potentially other variables that are statistically significant in a logistic regression model, they would not appreciably improve prediction and consequently were not included in the model. For example, alcohol consumption was a statistically significant variable when added to the multivariate logistic regression model ($P=.01$) but has little effect on the AUC.

We also tested the various components of the algorithm individually and in combinations that included or excluded age. These results are shown in TABLE 3. This demonstrates that age alone is clearly the best predictor of hip fracture, but added value is gained by the addition of other factors.

The ROC curve in FIGURE 2 shows the accuracy at different estimations of risk tested in all WHI trial participants. This shows the sensitivity and 1-specificity of the prediction of 5-year hip fracture risk for women at different levels of predicted risk. By application of this information, thresholds for further screening can be set based on acceptable risk and desire for certainty. For example, identifying women at risk using a threshold of a 1% 5-year risk would yield a true-positive rate (sensitivity) of about 50%, half of women who would have hip fractures within 5 years, but there would be a false-positive rate (1-specificity) of 15%. Half of the women who would have hip fractures in the next 5 years would be in this group, and 15% who were predicted to have hip fractures would not. A less stringent risk threshold of 0.5% would identify

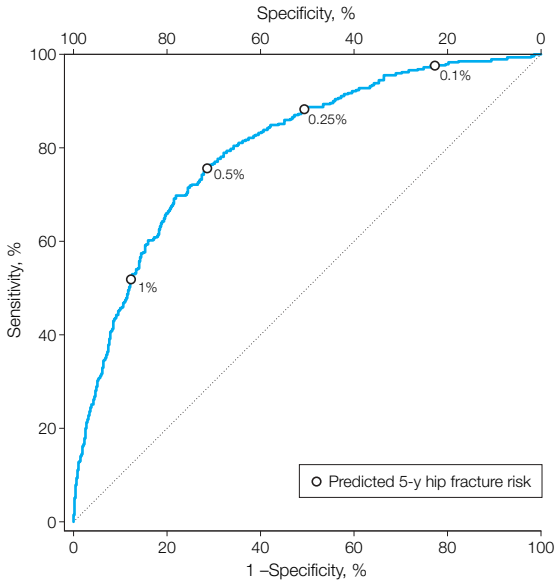
more women at risk of hip fracture, approximately 75% of women who would have hip fractures in the next 5 years, but one would double the rate of overdiagnosis. There would be a false-positive rate of 30%. Seventy-five percent of the hip fractures in the next 5 years would occur in this group, but 30% of the women predicted to have hip fractures would not.

As a final step, we compared the predictive value of the algorithm with the DXA measurements of the women whose BMD was measured. There were 10 750 women with DXA measurements who either completed at least 5 years of follow-up or experienced a hip fracture. The combined group had 80 hip fractures over a mean (SD) of 8.7 (1.2) years of follow-up; thus, the power to show a difference was small. ROC curves were calculated for the algorithm, the DXA, or a combination of the 2. These are shown in FIGURE 3. There was no statistically significant difference in the AUCs.

To show the relative utility of the models prediction based on the 5% highest-risk group for the DXA (T score ≤–2.5) and the 5% highest-risk group for the point-scoring method (score ≥21 points) were compared (A T score is a standard deviation of the bone density above the peak average for a young woman); 3.8% of the high-risk DXA group went on to have hip fractures in 5 years compared with 3.1% of the high-risk point-score group (TABLE 4).

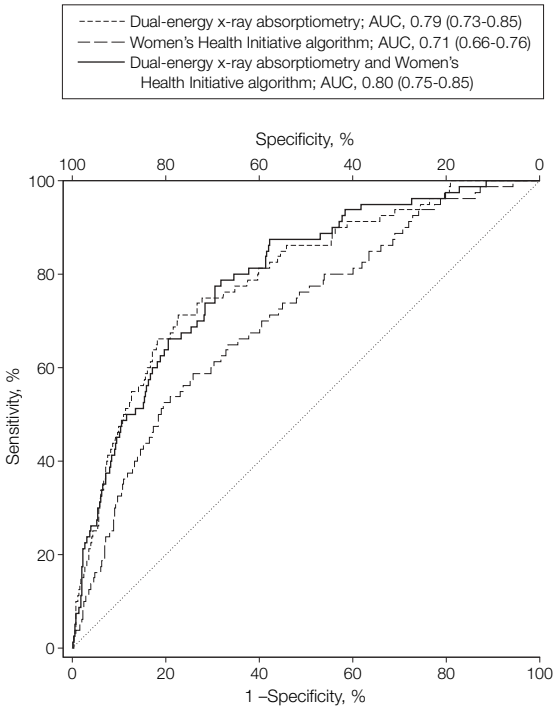
Noting that the number of women with a hip fracture in the subset with DXA scans was small, we compared the women identified by DXA and the point-scoring algorithm using the same cut points The results are shown in TABLE 5, which demonstrates the discordance between the 2 methods of prediction and actual outcomes. We also note that the women who were identified by only 1 method of prediction have had a substantial increased risk of fracture compared with the women who were not identified by either method.

Figure 2. Sensitivity and 1-Specificity of Receiver Operating Characteristic at Selected Percentage Predictions of 5-Year Risk of Hip Fracture



AUC indicates area under the curve. Blue curves in Figure 1 and Figure 2 are the same and are derived from the entire clinical trial cohort.

Figure 3. Comparison of the Women's Health Initiative Algorithm With Results From Dual-Energy X-ray Absorptiometry Scans



AUC indicates area under the curve. Data are based on a subset of 10 750 women with bone mass density measurements.

TABLE 6 shows the median 5-year hip fracture risk and the 2.5% upper and lower limits for this prediction, according to a simple sum point score that approximates the WHI probability score. This can also be done using the online algorithm.

Table 4. Comparison of the Prediction of 5-Year Risk of Hip Fracture in the 5% Highest Risk Group by Point Score and DXA

	Patients With Hip Fractures		Patients Without Hip Fractures	
	Observed	Predicted	Observed	Predicted
DXA				
>T score –2.5	60	57.0	10 165	10 169
<T score –2.5	20	23.0	504	501
WHI				
<21 points	65	64.9	10 203	10 203.1
≥21 Points	15	15.1	467	466.9

Abbreviations: DXA, dual-energy x-ray absorptiometry; T score, a standard deviation of bone density above the peak average for a young woman; WHI, Women's Health Initiative.

Table 5. Cross-tabulation of Women With Osteoporosis Identified by DXA and Women's Health Initiative Algorithm

	No. of Women			No. (%) of Women With Hip Fracture Within 5 Years	
DXA	WHI Score <21	WHI Score ≥21	Total	WHI Score <21	WHI Score ≥21
T score ≥–2.5	9859	367	**10 226**	50 (0.51)	10 (2.72)
T score <–2.5	409	115	**524**	15 (3.67)	5 (4.35)
Total	**10 268**	**482**	**10 750**		

Abbreviations: DXA, dual-energy x-ray absorptiometry; T score, a standard deviation of the bone density above the peak average for a young woman; WHI, Women's Health Initiative.

Table 6. Women's Health Initiative Estimated Probability of Hip Fracture Within 5 Years by Women's Health Initiative Hand Score

	WHI Probability Score, %[a]		
WHI Hand Score[b]	2.5th Percentile	Median (50th Percentile)	97.5th Percentile
≤7	<0.1	<0.1	<0.1
8	<0.1	<0.1	0.1
9	<0.1	0.1	0.2
10	<0.1	0.1	0.2
11	<0.1	0.2	0.3
12	0.1	0.2	0.4
13	0.1	0.3	0.5
14	0.2	0.4	0.6
15	0.2	0.5	0.8
16	0.3	0.6	1.0
17	0.4	0.8	1.3
18	0.5	1.0	1.6
19	0.7	1.3	2.0
20	0.9	1.6	2.6
21	1.1	2.1	3.3
22	1.4	2.7	4.3
23	2.0	3.5	>5.0
24	2.8	4.6	>5.0
≥25	3.6	>5.0	>5.0

Abbreviation: WHI, Women's Health Initiative.
[a] WHI probability of hip fracture within 5 years based on WHI observational study.
[b] Simple sum score approximating WHI probability score.

COMMENT

The large sample size, multiethnic composition, geographically diverse, ambulatory population, and adjudicated hip fracture outcomes in WHI has made it possible to develop a comprehensive model to predict the 5-year risk of hip fracture in postmenopausal women. Because we were working from one dataset, with 93 676 postmenopausal women to develop the model and more than 60 000 women to validate it, our conclusions are robust. Because of the great uniformity in the collection methods and uniformity in the factors included in the model creation and validation testing, plus significant differences in the frequency of risk factors, the model appears to be generalizable.

Instead of splitting the sample to have a training set and a test set, we were able to take advantage of the multiple components of WHI, using one group to develop the model, the training set, and another to validate it, the validation set. By including minority women in the model, predicting fracture risk extends risk factors for nonwhite women beyond race or ethnic background.

Age is a known major risk factor for fracture and continues to be the most powerful predictor of fracture risk, but we have demonstrated that the addition of a few readily available items of clinical information can enhance this prediction. As with many prediction models, one is faced with a trade-off between specificity and sensitivity. As can be seen from Table 4, most fractures occur in women who are predicted to be a low risk. Figure 2 clearly demonstrates this trade off.

The comparison of the DXA prediction with the algorithm is limited by sample size, and there is no statistically significant difference even though within these limits, the DXA appears to give marginally better results at an obviously greater cost. But before the algorithm is considered definitive, these 2 methods should be tested in other large populations. The role of each needs to be clarified relative to screening and treatment.

There are a several limitations to this study. Certain data, such as accurate classification of arthritis type or an objective measure of physical activity were not available, but those factors that were available were clearly defined. Dual-energy x-ray absorptiometry data were not available for all participants; however, this model may provide a low-cost general screening prediction model that has certain advantages for general use. In addition, a longer period than 5 years might provide more information for long-term fracture prediction. Unfortunately, 5 years is the maximum follow-up available across the components of the study at this time. Other prediction models, eg, the Gail model for breast cancer, also has been predictive of breast cancer over more than 5 years. The cohort continues to be followed up and future assessment of longer-term prediction models may be available in the future.

Some aspects of the study may limit its generalizability to other populations. Of note, the annual hip fracture rate in women older than 65 years estimated from the 2004 National Hospital Discharge Survey is 57/10 000 women compared with rates of 30 in the observational and 32 in the clinical trial of women participating in the WHI who were in the same age group. This may reflect the higher BMI in the WHI population, the truncation of upper age groups and the healthy volunteer effect. In addition, the national data includes institutionalized women who have much higher fracture rates. Women in the training set and the validation set differ in many ways: for example, the validation set included more than 27 000 women in the hormone therapy clinical trials. Hormone therapy is known to greatly influence fracture risk, but hormone therapy was not selected as a useful predictor in the observational study training data. The model from the training data still appeared valid in the validation data. The lower risk of fracture in the WHI population may or may not be corrected by the factors in the model. The answer will only come when the model is tested in disparate populations.

This study does not indicate whether women defined by the WHI algorithm to be at risk would benefit from measures to prevent hip fracture in contrast to those trials that have used DEXA-T scores as a criterion for treatment. Some women who would be classified as high risk (point score >21) in our study did not have low T scores, which is currently used as the gold standard for defining osteoporosis. Further studies are needed to define the clinical implications of this algorithm and to confirm treatment benefits for those delineated by the WHI risk classification to be an increased risk for hip fracture. Ultimately, the decision of whom to further screen for osteoporosis and whom to treat will need to be based on available resources and major social and political judgments. Knowing the 5-year risk of fracture will permit patients and physicians to make informed choices when balancing making lifestyle changes against undergoing medical interventions. Publication of these results, along with the user-friendly tool for their application, will permit others to rapidly test their utility. However, we believe 11 readily available clinical variables offer a simple means of stratifying the 5-year risk of hip fracture in postmenopausal women.

Author Contributions: Dr Robbins 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: Robbins, Kooperberg, Watts, Wactawski-Wende, Lewis, Chen, LeBoff.
Acquisition of data: Robbins, Kooperberg, Watts, Wactawski-Wende, Jackson, Lewis, Stefanick.
Analysis and interpretation of data: Aragaki, Kooperberg, Watts, Wactawski-Wende, Jackson, Chen, LeBoff, Stefanick, Cauley.
Drafting of the manuscript: Robbins, Aragaki, Kooperberg.
Critical revision of the manuscript for important intellectual content: Robbins, Aragaki, Kooperberg, Watts, Wactawski-Wende, Jackson, Lewis, Chen, LeBoff, Stefanick, Cauley.
Statistical analysis: Aragaki, Kooperberg, LeBoff.
Obtained funding: Robbins, Wactawski-Wende, Lewis, Stefanick.
Administrative, technical, or material support: Robbins, Kooperberg, Wactawski-Wende, Jackson, Lewis, LeBoff, Cauley.
Study supervision: Robbins, Kooperberg, Watts, Wactawski-Wende.

Financial Disclosures: Dr Cauley reports that she has received research support from Merck & Co, Eli Lilly & Co, Pfizer Pharmaceuticals, and Novartis Pharmaceuticals; has served as a consultant for Eli Lily and Novartis; and serves on the Merck speaker's bureau. No other financial conflicts were disclosed.

WHI Investigators: Program Office: National Heart, Lung, and Blood Institute, Elizabeth Nabel, Jacques Rossouw, Shari Ludlam, Linda Pottern, Joan McGowan, Leslie Ford, and Nancy Geller.
Clinical Coordinating Center: *Fred Hutchinson Cancer Research Center, Seattle, Wash:* Ross Prentice, Garnet Anderson, Andrea LaCroix, Charles L. Kooperberg, Ruth E. Patterson, Anne McTiernan; *Wake Forest University School of Medicine, Winston-Salem, North Carolina:* Sally Shumaker; *Medical Research Labs, Highland Heights, Kentucky:* Evan Stein; *University of California at San Francisco, San Francisco:* Steven Cummings.
Clinical Centers: *Albert Einstein College of Medicine, Bronx, New York:* Sylvia Wassertheil-Smoller; *Baylor College of Medicine, Houston, Texas:* Jennifer Hays; *Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts:* JoAnn Manson; *Brown University, Providence, Rhode Island:* Annlouise R. Assaf; *Emory University, Atlanta, Georgia:* Lawrence Phillips; *Fred Hutchinson Cancer Research Center, Seattle, Washington:* Shirley Beresford; *George Washington University Medical Center, Washington, DC:* Judith Hsia; *Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, California:* Rowan Chlebowski; *Kaiser Permanente Center for Health Research, Portland, Ore:* Evelyn Whitlock; *Kaiser Permanente Division of Research, Oakland, Calif:* Bette Caan; *Medical College of Wisconsin, Milwaukee:* Jane Morley Kotchen; *MedStar Research Institute/Howard University, Washington, DC:* Barbara V. Howard; *Northwestern University, Chicago/Evanston, Illinois:* Linda Van Horn; *Rush Medical Center, Chicago, Illinois:* Henry Black; *Stanford Prevention Research Center, Stanford, California:* Marcia L. Stefanick; *State University of New York at Stony Brook, Stony Brook:* Dorothy Lane; *The Ohio State University, Columbus:* Rebecca Jackson; *University of Alabama at Birmingham, Birmingham:* Cora E. Lewis; *University of Arizona, Tucson/Phoenix:* Tamsen Bassford; *University at Buffalo, Buffalo, New York:* Jean Wactawski-Wende; *University of California at Davis, Sacramento:* John Robbins; *University of California at Irvine:* F. Allan Hubbell; *University of California at Los Angeles, Los Angeles:* Howard Judd; *University of California at San Diego, LaJolla/Chula Vista:* Robert D. Langer; *University of Cincinnati, Cincinnati, Ohio:* Margery Gass; *University of Florida, Gainesville/Jacksonville:* Marian Limacher; *University of Hawaii, Honolulu:* David Curb; *University of Iowa, Iowa City/Davenport:* Robert Wallace; *University of Massachusetts/Fallon Clinic, Worcester:* Judith Ockene; *University of Medicine and Dentistry of New Jersey:* Norman Lasser; *University of Miami, Miami, Florida:* Mary Jo O'Sullivan; *University of Minnesota, Minneapolis:* Karen Margolis; *University of Nevada, Reno:* Robert Brunner; *University of North Carolina, Chapel Hill:* Gerardo Heiss; *University of Pittsburgh, Pittsburgh, Pennsylvania:* Lewis Kuller; *University of Tennessee, Memphis:* Karen C. Johnson; *University of Texas Health Science Center, San Antonio:* Robert Brzyski; *University of Wisconsin, Madison:* Gloria E. Sarto; *Wake Forest University School of Medicine, Winston-Salem, North Carolina:* Denise Bonds; and *Wayne State University School of Medicine/Hutzel Hospital, Detroit, Michigan:* Susan Hendrix.
Funding/Support: The Women's Health Initiative program was funded by the National Heart, Lung, and Blood Institute of the National Institutes of Health, Department of Health and Human Services.
Role of the Sponsor: The funding organization had representation on the steering committee, which gov-

erned the design and conduct of the study, the interpretation of the data, and the preparation and approval of manuscripts. The National Heart, Lung, and Blood Institute Program Office reviewed the manuscript prior to publication.

REFERENCES

1. National Center for Health Statistics. National Hospital Discharge and Ambulatory Surgery Data. http://www.cdc.gov/nchs/about/major/hdasd/nhdstab.htm. Accessed July 7, 2007.
2. Braithwaite RS, Col NF, Wong JB. Estimating hip fracture morbidity, mortality and costs. *J Am Geriatr Soc*. 2003;51(3):364-370.
3. Cummings SR, Nevitt MC, Browner WS, et al. Risk factors for hip fracture in white women: Study of Osteoporotic Fractures Research Group. *N Engl J Med*. 1995;332(12):767-773.
4. Black DM, Steinbuch M, Palermo L, et al. An assessment tool for predicting fracture risk in postmenopausal women. *Osteoporos Int*. 2001;12(7):519-528.
5. Buist DS, LaCroix AZ, Manfredonia D, Abbott T. Identifying postmenopausal women at high risk of fracture in populations: a comparison of three strategies. *J Am Geriatr Soc*. 2002;50(6):1031-1038.
6. McGrother CW, Donaldson MM, Clayton D, Abrams KR, Clarke M. Evaluation of a hip fracture risk score for assessing elderly women: the Melton Osteoporotic Fracture (MOF) study. *Osteoporos Int*. 2002;13(1):89-96.
7. van Staa TP, Leufkens HG, Cooper C. Utility of medical and drug history in fracture risk prediction among men and women. *Bone*. 2002;31(4):508-514.
8. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *Eur J Epidemiol*. 1991;7(4):403-422.
9. Wainwright SA, Marshall LM, Ensrud KE, et al. Hip fracture in women without osteoporosis. *J Clin Endocrinol Metab*. 2005;90(5):2787-2793.
10. Fang J, Freeman R, Jeganathan R, Alderman MH. Variations in hip fracture hospitalization rates among different race/ethnicity groups in New York City. *Ethn Dis*. 2004;14(2):280-284.
11. Ritenbaugh C, Patterson RE, Chlebowski RT, et al. The Women's Health Initiative Dietary Modification trial: overview and baseline characteristics of participants. *Ann Epidemiol*. 2003;13(9)(suppl):S87-S97.
12. Stefanick ML, Cochrane BB, Hsia J, Barad DH, Liu JH, Johnson SR. The Women's Health Initiative postmenopausal hormone trials: overview and baseline characteristics of participants. *Ann Epidemiol*. 2003;13(9)(suppl):S78-S86.
13. Jackson RD, LaCroix AZ, Cauley JA, McGowan J. The Women's Health Initiative calcium-vitamin D trial: overview and baseline characteristics of participants. *Ann Epidemiol*. 2003;13(9)(suppl):S98-S106.
14. Design of the Women's Health Initiative clinical trial and observational study: The Women's Health Initiative Study Group. *Control Clin Trials*. 1998;19(1):61-109.
15. Curb JD, McTiernan A, Heckbert SR, et al. Outcomes ascertainment and adjudication methods in the Women's Health Initiative. *Ann Epidemiol*. 2003;13(9)(suppl):S122-S128.
16. Anderson GL, Manson J, Wallace R, et al. Implementation of the Women's Health Initiative study design. *Ann Epidemiol*. 2003;13(9)(suppl):S5-S17.
17. Jackson RD, LaCroix AZ, Gass M, et al. Calcium plus vitamin D supplementation and the risk of fractures. *N Engl J Med*. 2006;354(7):669-683.
18. Anderson GL, Limacher M, Assaf AR, et al. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *JAMA*. 2004;291(14):1701-1712.
19. Cauley JA, Robbins J, Chen Z, et al. Effects of estrogen plus progestin on risk of fracture and bone mineral density: the Women's Health Initiative randomized trial. *JAMA*. 2003;290(13):1729-1738.
20. Langer RD, White E, Lewis CE, Kotchen JM, Hendrix SL, Trevisan M. The Women's Health Initiative Observational Study: baseline characteristics of participants and reliability of baseline measures. *Ann Epidemiol*. 2003;13(9)(suppl):S107-S121.
21. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol*. 2003;13(9)(suppl):S18-S77.
22. Chen Z, Kooperberg C, Pettinger MB, et al. Validity of self-report for fractures among a multiethnic cohort of postmenopausal women: results from the Women's Health Initiative observational study and clinical trials. *Menopause*. 2004;11(3):264-274.
23. Hsia J, Wu L, Allen C, et al. Physical activity and diabetes risk in postmenopausal women. *Am J Prev Med*. 2005;28(1):19-25.
24. Wolf AM, Hunter DJ, Colditz GA, et al. Reproducibility and validity of a self-administered physical activity questionnaire. *Int J Epidemiol*. 1994;23(5):991-999.
25. Ainsworth BE, Haskell WL, Leon AS, et al. Compendium of physical activities: classification of energy costs of human physical activities. *Med Sci Sports Exerc*. 1993;25(1):71-80.
26. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). *Am J Prev Med*. 1994;10(2):77-84.
27. Patterson RE, Kristal AR, Tinker LF, Carter RA, Bolton MP, Agurs-Collins T. Measurement characteristics of the Women's Health Initiative food frequency questionnaire. *Ann Epidemiol*. 1999;9(3):178-187.
28. Neuhouser ML, Patterson RE, King IB, Horner NK, Lampe JW. Selected nutritional biomarkers predict diet quality. *Public Health Nutr*. 2003;6(7):703-709.
29. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J R Stat Soc [Ser B]*. 1974;36(2):111-147.
30. Lloyd-Jones DM, Liu K, Tian L, Greenland P. Narrative review: assessment of C-reactive protein in risk prediction for cardiovascular disease. *Ann Intern Med*. 2006;145(1):35-42.
31. Kooperberg C. *polspline: Polynomial spline routines* [computer program]. R package version 1.0.14. 2007.
32. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940-3941.
33. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2007. http://www.R-project.org. Accessed November 6, 2007.

