The Validity of Health Risk Appraisals for Coronary Heart Disease: Results from a Randomized Field Trial

ABSTRACT

Background: While health risk appraisals (HRAs) are becoming increasingly popular as tools for health assessment and health education, comparatively little is known about the accuracy of these risk estimates.

Methods: A field trial among 732 randomly selected adults ages 25 to 65 years was conducted to assess the validity of the risk scores produced by four widely used HRAs.

Results: Self-reported HRA risk scores for cigarette smoking and relative weight were generally accurate, but correlations between physiological measurements and scores for blood pressure, cholesterol, and physical activity were always lower than .51. Correlations between epidemiologic estimates of the probability of CHD death and HRA total risk scores ranged from .13 to .75; partial correlations adjusting for age, race, and gender ranged from .12 to .47.

Conclusions: The HRAs chosen for the field trial exhibited modest correlations with the CHD mortality risk predicted by the epidemiologic model. Mathematical errors made by respondents completing self-scored instruments and lack of knowledge regarding physiologic status decrease the accuracy of HRA risk estimates. (*Am J Public Health* 1991;81:466–470)

Kevin W. Smith, MA, Sonja M. McKinlay, PhD, and John B. McKinlay, PhD

Introduction

Health risk appraisals (HRAs) are designed to estimate a person's mortality or morbidity risk for various diseases based on such characteristics as medical history, blood pressure, and smoking habits. HRAs have become increasingly popular as health education tools. Recent estimates indicate that more than 200 organizations in the United States currently use HRAs to promote health awareness, and that between five and 15 million people have completed one of these instruments.[1] HRAs are also frequently included in work-site health promotion activities.[2] Despite the burgeoning popularity of HRAs, comparatively little is known about the validity of the risk estimates produced by them. Several recent reviews of HRA usage have stressed the need for further research on the accuracy of the predictions made by these instruments.[3–5]

In an earlier paper,[6] we examined the accuracy of the scoring systems employed by 41 different HRAs to estimate coronary heart disease (CHD) mortality risk. Correlations between HRA risk scores and epidemiologic estimates were found to be influenced by the sophistication of the HRA scoring system, the number of risk factors included in the instrument, the range of categories available for each risk factor, and the sensitivity of the scoring system to the effects of age and gender. Foxman and Edington[7] have also examined HRA accuracy with respect to mortality from all causes. These two studies based their assessments of HRA accuracy on physical examinations in the Framingham Heart Study and Tecumseh Community Health Study, respectively. Since the respondents who complete HRAs may or may not be aware of the appropriate values for risk factors, the use of actual physiological measurements in these studies is likely to have overstated the accuracy that would be obtained from self-reported risk scores.

This paper examines HRA validity based on self-reported data from a general population. Three questions are addressed: 1) How accurate are respondents' self-reports for HRA risk factors? 2) How accurate are HRA estimates of the risk of CHD mortality based on these self-reports? 3) To what extent is the validity of HRA risk scores affected by reporting errors and computational mistakes made by respondents?

Methods

A field trial of health risk appraisal instruments was conducted by the Cambridge Research Center of the American Institutes for Research in 1987. The trial was specifically designed to assess the reliability and validity of four representative HRAs in a general population. Respondents between the ages of 25 and 65 were identified through random digit dialing of 15 telephone exchanges in the Boston metropolitan area. Sampling and interviewing procedures for the field trial have been described elsewhere.[8]

During home visits, trained field technicians administered a standardized per-

From the Cambridge Research Center, American Institutes for Research, Watertown, MA. Address reprint requests to Kevin W. Smith, Research Scientist, New England Research Institute, 9 Galen Street, Watertown, MA 02172. Dr. S. McKinlay is President and Dr. J. McKinlay is Vice President of NERI.

This paper, submitted to the Journal March 3, 1989, was revised and accepted for publication August 13, 1990.

TABLE 1—Risk Score Ranges by HRA

	Arizona Heart Institute	CDC	Determine Your Medical Age	RISKO
Heart Attack Risk	0–50	2–21,969	—	0–27
Appraised Age	—	—	12.0–70.6	—
Smoking	0–10	.8–2.1	−6–24	−2–4
Relative Weight	0–3	.8–1.6	−6–24	−2–3
Blood Pressure	0–6	.4–1.1	0–12	−5–6
Cholesterol	—	.5–1.6	0–6	−3–3
Physical Activity	0–3	.6–2.5	−12–12	—

TABLE 2—Validation Measures for Risk Items and for Total Risk Scores

Self-Reported Risk Factors	Validation Measure
Blood Pressure	2nd Sphygmomanometer Reading (mmHg)
Total Serum Cholesterol	Laboratory Lipid Analysis (mg/dl)
Physical Activity	Harvard Alumni Activity Survey Scale (natural log of kcal expenditure for previous week)
Relative Weight	Quetelet Index (kg/m^2)
Smoking	Cigarettes per Day
Heart Attack Risk Appraised Age	NHANES I Epidemiologic Followup: CHD mortality risk estimate

sonal interview and one of the four HRAs to be evaluated by the project. Technicians were instructed not to help respondents complete HRA items or compute risk scores. The technician also measured blood pressure, height, and weight, and drew a venous blood sample for cholesterol analysis following the protocol established by the Centers for Disease Control (CDC). In-home physiological measurements have been shown to be as reliable and valid as those performed in a central location using standard equipment.[9]

Four widely used HRAs were chosen to represent the major types of computer-scored and self-scored instruments available in 1987: 1) Health Risk Appraisal (Centers for Disease Control), 2) The Heart Test (Arizona Heart Institute), 3) RISKO (American Heart Association), and 4) Determine Your Medical Age (Blue Cross/Blue Shield of New York). These were found to be among the most valid instruments within major HRA classifications in our previous work.[6] Since the field trial, the CDC instrument has been replaced by the Carter Center's "Healthier People" HRA.[10] The Carter Center's questionnaire and reporting format are similar to those of the earlier CDC version, but it uses an updated algorithm for assessing risk.

Each instrument requires a respondent to select categories that best describe personal habits, physiological status, and medical history. Point values or weights associated with these categories are then combined to produce an overall estimate of risk. In these four HRAs, total risk is summarized either by a heart-attack risk scale, with higher values representing greater risk, or by appraised age. Appraised age is the age at which the average mortality risk for all adults of the same gender and race is equal to the risk projected for a particular individual. If a person's appraised age is higher than his chronological age, that person has a greater than average chance of dying. The range of total risk scores and point values or weights for five major CHD risk items is shown in Table 1.

The CDC HRA, which requires computer processing, provides probability estimates for the 12 leading causes of death based on the Geller-Gesner method.[11] The Arizona Heart Institute's Heart Test and RISKO are self-scored instruments focusing on heart attack risk. In Determine Your Medical Age, appraised age is calculated by adjusting an individual's chronological age for medical history and lifestyle factors.

A standard psychometric definition for validity was adopted for this study; that is, that validity refers to the degree to which an instrument measures what it is intended to measure.[12,13] In order to assess the validity of a measure, some criterion or "gold standard" must be available for comparison. Table 2 shows the validation measures used in this study. Physiological values served as standards for three of the five CHD risk items. The standards for physical activity (kilocalorie expenditure for the previous week as calculated from the Harvard Alumni Activity Survey Scale)[14] and cigarette smoking were based on self-reports collected during personal interviews. HRA scores for heart attack risk or appraised age were compared to estimates from the NHANES I Epidemiologic Followup Study (NHEFS).[15] Separate NHEFS logistic models for men and women provide estimates of the probability of death due to coronary heart disease over a ten-year period based on a large (N = 4,013) representative sample of white adults.

Several statistical techniques can be employed to assess the extent of agreement or conformity between two measurements, including sensitivity, specificity, correlations, mean differences, and standard deviations for differences. Which technique is most appropriate remains a controversial issue.[16,17] However, the self-reported risk scores and the validation measures in this study cannot be directly compared because they were not measured on the same scale. Therefore, product-moment correlations between risk scores and validation measures were employed as the primary indicator of accuracy.[12]

As the criterion measure for CHD risk, the 10-year heart disease mortality probability for each respondent was calculated using the gender-specific NHEFS models. Physiological measures taken by field technicians, rather than self-reports, were used to make the logistic estimates as accurate as possible. According to the logistic models, a few respondents had mortality risks that were much higher than the risks for the great majority of cases. To prevent these outliers from distorting the correlation coefficients, the natural logarithm of the 10-year mortality risk per 100,000 persons was used in the analyses.

Three HRA total risk scores were computed for each respondent: 1) the score reported by the respondent, 2) the risk score corrected for mathematical errors, and 3) the score that would have been obtained on the basis of the blood pressure, cholesterol, height, and weight measurements taken by field technicians. The first of these scores does not apply to the CDC instrument, which is calculated by computer. Each of these three risk scores was then correlated with the log-

TABLE 3—Demographic Characteristics of Field Trial Participants and Nonparticipants*

Characteristics	Participants: Physiologic Status Measured at Baseline (N = 309)	Participants: Physiologic Status Measured at Follow-up (N = 423)	Nonparticipants (n = 544)
Age (years)	39.8 (10.7)	38.9 (10.6)	40.2 (11.1)
Percent female	59.9 (49.1)	56.0 (49.7)	60.5 (48.9)
Percent Black/Hispanic	11.0 (31.4)	6.7 (24.9)	8.2 (27.4)
Income (thousands)†	41.4 (11.3)	41.0 (10.3)	40.5 (10.1)
Self-reported health status**	1.94 (.87)	2.04 (.94)	2.06 (.91)
Worry about health††	2.56 (.84)	2.53 (.84)	2.52 (.84)

*Mean and (standard deviation)
†Mean household income in subject's telephone exchange
**Five-point scale; 1 = excellent, 5 = poor
††Four-point scale; 1 = not at all, 4 = most of the time

transformed probability estimates derived from the NHEFS logistic model. Logarithmic transformations were also applied to the CDC probabilities. CHD mortality rates increase monotonically with age and are considerably higher for men than women. To remove these effects from the validity assessments, partial correlations were also computed for each HRA score, adjusting for age, gender, and race.

Results

Baseline interviews were completed by 732 of 1,276 eligible adults (57.4 percent) contacted for the field trial. A logistic regression analysis of response rates indicated that the probability of obtaining a completed interview was not affected by demographic characteristics (gender, race, age, and average income level in the telephone exchange), self-reported health status, or concern about the state of one's health. Similarities between field trial participants and eligible subjects who declined to be interviewed are shown in Table 3. The median age of the sample was 38 years. Respondents were somewhat younger and healthier than the general population because persons with diabetes, hypertension, or evidence of overt cardiovascular disease were screened out of the sample.

Validity of CHD Risk Factors

The first set of analyses focused on five specific CHD risk items. These analyses were restricted to respondents whose physiological measurements were completed during the baseline interview. Table 4 shows the correlations between self-reported risk scores and validation measures for each of the four HRAs. These results reveal some sharp distinctions among the individual risk factors. The coefficients for cigarette smoking and relative weight were always .6 or greater for each of the four instruments. On the other hand, none of the correlations for physical activity, blood pressure, or cholesterol were higher than .5. The HRA risk scores for these items accounted for less than 25 percent of the variance in the physiological measures. The smallest correlations were found for blood pressure and cholesterol. This occurs primarily because many people do not know the appropriate values and either assume that their levels are average or make inaccurate guesses. The magnitudes of these correlations were relatively consistent from HRA to HRA; there was more variation among different risk factors than among instruments.

Validity of Heart Attack Risk Scores and Appraised Age

The second set of analyses examined the validity of HRA scores for heart attack risk and appraised age. There are two potential sources of error in total risk scores. First, respondents may be unaware of the appropriate values for physiological measures. When asked to specify exact values during in-person interviews, only 27 percent of the respondents gave systolic blood pressure values within 10 mmHg of those recorded by the field technicians, and only 4 percent estimated their total serum cholesterol to within 20 mg/dl of laboratory readings.

For self-scored appraisal instruments, a second source of error lies in computational mistakes made by respondents when calculating their total risk scores. Many of the field trial participants had difficulty computing these scores (Table 5). Error rates were affected both by the mathematical skills required by each instrument (addition, subtraction, or division) and by respondents' educational attainment (persons without four-year college degrees were twice as likely to make mistakes as those with baccalaureate or advanced degrees).

To assess the effects of these potential sources of error, separate correlations were computed for the risk scores reported by respondents, scores corrected for mathematical errors, and the scores that would have resulted from the physiological measurements taken by field technicians. The last of these three scores is similar to estimates based on physical examination data evaluated in our previous study.[6]

The upper panel of Table 6 shows the correlations for each HRA based on all respondents who were able to calculate risk scores and had a complete set of physiological measurements. The CDC risk appraisal had the highest correlations. This computerized instrument uses a more sophisticated scoring algorithm than the other HRAs and was the only one of the four that, like the logistic models, provides probability estimates for CHD mortality over a ten-year period. The CDC estimates, however, were consistently higher than the risks predicted by the NHEFS models (average risks of 1,522 per 100,000 compared with 1,173 per 100,000). Determine Your Medical Age had the next highest set of correlations, even though this instrument is not designed specifically for heart disease. Mathematical errors played an important role in the validity of this HRA, as the average coefficient increased from .41 to .55 when respondents' mistakes were corrected. Determine Your Medical Age also benefits considerably from the fact that CHD risk increases greatly with age, so that any HRA based on appraised age should have a reasonably high correlation in this type of analysis.

The pattern of results was considerably different for the Arizona Heart Institute HRA than for the other instruments. The correlations for this HRA were always about .52, regardless of whether computational errors were corrected or actual physiological values were substituted for the respondent's choices. This probably occurred because comparatively few people completing this instrument made errors in their risk calculations and because cholesterol values are not re-

TABLE 4—Correlations (and 95 Percent Confidence Intervals) between Validation Measures and Risk Item Scores

Risk Items	Arizona Heart Institute (N = 67–68) r	95% CI	CDC (N = 63–77) r	95% CI	Determine Your Medical Age (N = 64–70) r	95% CI	RISKO (N = 55–64) r	95% CI
Cigarettes Smoked per Day	.85	(.77–.91)	.66	(.52–.77)	.75	(.62–.84)	.80	(.69–.88)
Relative Weight	.74	(.61–.84)	.86	(.79–.91)	.64	(.47–.76)	.70	(.55–.81)
Physical Activity	–.27	(–.48 to –.04)	–.48	(–.64 to –.28)	–.47	(–.64 to –.26)	—	—
Systolic Blood Pressure	.47	(.26–.64)	.36	(.14–.54)	.00	(–.23–.24)	.50	(.29–.67)
Total Serum Cholesterol	—	—	.40	(.17–.59)	.18	(–.06–.41)	.26	(.00–.49)

TABLE 5—Percentage of Baseline Respondents Making Computational Errors When Calculating HRA Risk Scores

	Arizona Heart Institute (N = 157)	Determine Your Medical Age (N = 170)	RISKO (N = 155)
Score computed correctly	86.6	41.8	74.8
Score computed incorrectly	7.0	32.9	14.2
Could not compute score	6.4	25.3	11.0
Totals	100%	100%	100%

TABLE 6—Correlations between HRA Risk Scores and Natural Logarithm of NHEFS Ten-year CHD Mortality Risk Estimates

	Arizona Heart Institute (N = 124)	CDC (N = 137)	Determine Your Medical Age (N = 104)	RISKO (N = 112)
Zero-order correlations:				
Self-reported	.52	—	.41	.13
Corrected for errors	.52	.75	.55	.09
Physiologic field measures	.52	.77	.56	.40
R^2 for NHEFS estimate regressed on sex, age, and race	.85	.90	.87	.86
Partial correlations*:				
Self-reported	.42	—	.12	.19
Corrected for errors	.39	.47	.42	.39
Physiologic field measures	.39	.51	.43	.68

*Adjusted for respondent's gender, age, and race.

quired. RISKO had the smallest correlation with logistic estimates for the scores reported by respondents (.13), but in contrast to the other three HRAs, RISKO scores represent relative rather than absolute mortality risk.

A more refined assessment of validity can be made by examining partial correlations between HRA scores and the logistic estimates after controlling for age, gender, and race, as shown in the lower panel of Table 6. These three demographic characteristics consistently explained 85 to 90 percent of the variation in the log-transformed logistic rates. The resulting partial correlations were smaller than the unadjusted correlations in Table 5 for all HRAs except RISKO. The partials for self-reported Medical Age and RISKO scores were less than .2, but these coefficients more than doubled when mathematical errors were corrected. The magnitudes of the partial correlations also showed small improvements for three HRAs and a large increase for RISKO when physiological measurements taken in the field were substituted for the values selected by respondents.

Discussion

The results of this randomized field trial have several important implications for the accuracy of HRA instruments when used by a general population.

First, self-reported risk scores for cigarette smoking and relative weight appear to be accurate enough for use in HRAs. However, reports for physical activity, blood pressure, and cholesterol were frequently inaccurate when compared with more objective measures. Many people don't know their blood pressure or cholesterol levels and don't do a very good job of guessing what these values are. In the field trial, 27 percent of the sample reported systolic blood pressure values that were within 10 mmHg of the readings taken by field technicians and only 4 percent gave cholesterol values that were within 20 mg/dl.

The HRAs chosen for the field trial exhibited modest correlations with the CHD mortality risk estimates predicted by the NHEFS epidemiologic model. After adjusting for variations attributable to age, gender, and race, the partial correlations for all three of the self-scored instruments were approximately .4 when computational errors had been corrected. The partial correlation for the CDC HRA, the only instrument that produced probability estimates, was only slightly higher (.47). Moreover, like other HRAs based on the Geller-Gesner techniques, CDC's ten-year mortality risk predictions were consistently higher than those produced by the NHEFS data.[6,7] Several of the more recently developed HRAs incorporate logistic equations that may be expected to yield higher partial correlations than the field trial instruments.

The trial results also indicate that, when used in the field, the accuracy of the self-scored HRAs is attenuated by mathematical errors made by respondents and, to a lesser extent, by a lack of knowledge of the exact values for blood pressure, cholesterol, height, and weight. Computational errors substantially reduced the correlations for RISKO and Determine Your Medical Age, but not for the Arizona

Heart Institute instrument. This appears to be a function of the computational burden a self-scored HRA imposes on respondents. Determine Your Medical Age requires three mathematical operations (addition, subtraction, and division); RISKO, two (addition and subtraction), while the Arizona Heart Institute scores involve only addition. Computational errors and uncertainty regarding one's physiological status have also been found to lower the test-retest reliability of these HRAs.[8] Providing respondents with accurate physiological measurements and checking HRA calculations for errors will therefore improve both the reliability and the validity of self-scored risk estimates.

Finally, it should be noted that the analyses presented in this paper are restricted to CHD mortality. Since the risk factors for CHD are relatively well established, HRA risk estimates may be less accurate for other, less frequent causes of death and morbidity. □

Acknowledgments

This project was supported by Grant No. HL32141 from the National Heart, Lung, and Blood Institute, National Institutes of Health, Co-Principal Investigators Drs. John B. McKinlay and Sonja M. McKinlay. An earlier version of this paper was presented at the 116th Annual Meeting of the American Public Health Association, Boston, Massachusetts, November 17, 1988. The authors would like to thank Dr. Jesse Berlin and Patricia Foxen for their helpful comments on this manuscript.

References

1. Schoenbach VJ: Appraising health risk appraisal (editorial). Am J Public Health 1987; 77:409–411.
2. Fielding JE, Piserchia PV: Frequency of worksite health promotion activities. Am J Public Health 1989; 79:16–20.
3. Beery W, Schoenbach VJ, Wagner EH, *et al:* Health Risk Appraisal: Methods and Programs, with Annotated Bibliography. National Center for Health Services Research and Health Care Technology Assessment, DHHS Pub. No.(PHS)86-3396. Washington, DC: Govt Printing Office, June 1986.
4. Fielding JE: Appraising the health of health risk appraisal (editorial). Am J Public Health 1982; 72:337–340.
5. DeFriese GH: A research agenda for personal health risk assessment. Health Services Research 1987; 22:581–594.
6. Smith KW, McKinlay SM, Thorington BD: The validity of health risk appraisal instruments for assessing coronary heart disease risk. Am J Public Health 1987; 77:419–424.
7. Foxman B, Edington DW: The accuracy of health risk appraisal in predicting mortality. Am J Public Health 1987; 77:971–974.
8. Smith KW, McKinlay SM, McKinlay JB: The reliability of health risk appraisals: A field trial of four instruments. Am J Public Health 1989; 79:1603–1607.
9. McKinlay SM, Kipp DM, Johnson P, Downey K, Carleton RA: A field approach for obtaining physiological measures in surveys of general populations: Response rates, reliability, and costs. *In:* Health Survey Research Methods: Proceedings of the Fourth Conference on Health Survey Research Methods. DHHS Pub. No. (PHS)84-3346, 1984; 195–204.
10. Hutchins EB: Healthier People: The Carter Center of Emory University Health Risk Appraisal Program, Vol 4. Atlanta, GA: Emory University, 1988.
11. Hall JH, Zwemer JD: Prospective Medicine. Indianapolis: Methodist Hospital of Indiana, 1979.
12. Nunnally JC: Psychometric Theory. New York: McGraw-Hill, 1967.
13. Carmines EG, Zeller RA: Reliability and Validity Assessment. Beverly Hills, CA: Sage, 1979.
14. Paffenbarger RS, Hyde RT, Wing AL, *et al:* Physical activity, all-cause mortality, and longevity of college alumni. New Engl J Med 1986; 314:605–613.
15. Leaverton PE, Sorlie PD, Kleinman JC, *et al:* Representativeness of the Framingham risk model for coronary heart disease mortality: A comparison with a national cohort study. J Chron Dis 1987; 40:775–784.
16. Kramer MS, Feinstein AR: Clinical biostatistics: The biostatistics of concordance. Clin Pharmacol Ther 1981; 29:111–123.
17. Maclure M, Willett WC: Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987; 126:161–169.