

MONOGRAPH

KEY WORDS: Multivariate Prediction, Surrogate Endpoint, Outcomes Research, Pharmacoeconomics, Disease Modeling, Clinical Trial Modeling, Human Diagnostics, Disease Prevention, Predictive Medicine

Synthesis analysis: Innovative biostatistics for drug development and human diagnostics

BioSignia, Inc.

The ability to predict multifactor chronic disease morbidity and mortality endpoints with biostatistics could play a more meaningful role in drug development and human diagnostics if there was a practical method for designing and constructing predictive equations. For example, multivariate regression represents a powerful tool for predicting clinical endpoints, however, the only valid approach for constructing these equations is from empirical data collected during lengthy longitudinal studies or clinical trials. Three limitations of this approach diminish its practicality: (1) the entire set of independent variables to be correlated with a clinical endpoint must be defined *a priori* to executing the longitudinal research; (2) data collection typically is greater than five years; (3) disparate sets of variables correlated with a common endpoint cannot be combined, i.e., the entire set of independent variables must be studied *in toto* looking for association with a common clinical endpoint.

Synthesis analysis is a new method for constructing multivariate predictive equations. This method involves defining a set of independent variables and respective

clinical endpoint, and then synthesizing a predictive equation from univariate correlation data presented in completed longitudinal or clinical research studies. Unlike meta-analysis, synthesis analysis allows disparate sets of variables that have been correlated with a common endpoint to be combined into a single multivariate predictive equation. Thus, defining a set of independent variables for correlation with a clinical endpoint can take place *a posteriori* to conducting lengthy longitudinal research. As such, synthesis analysis brings a level of practicality to constructing multivariate predictive equations that may allow for a more central role for predictive statistics in the drug development and human diagnostics industries.

Synthesis analysis statistics may produce new tools for modeling drug efficacy and new approaches for developing surrogate endpoints, which ultimately will support pharmaceutical and biotech companies' initiatives to bring products to market sooner. Healthcare economists and clinical outcomes researchers can use this statistical method to model and conduct complex pharmacoeconomic research in relatively short timeframes.

Researchers in the field of human diagnostics, disease prevention, and personalized predictive medicine can utilize synthesis analysis to rapidly incorporate new genomic, proteomic, and biochemical findings to existing state-of-the-art prognostic and predictive tests for disease onset and mortality. Thus, synthesis analysis can provide the means for continuously strengthening the discriminatory power of diagnostic tools for assessing disease onset and disease complications.

In this paper, we describe the synthesis analysis technique for constructing multivariate predictive equations from disparate longitudinal research. We also present findings on the validity of the methodology and discuss practical applications in context of drug development and human diagnostics.

PREDICTIVE STATISTICS

The real challenge of developing multivariate regression equations for specific disease endpoints is the restriction of statistical methods and the limitation of available data. Most prediction models have been developed directly from empirical data collected during longitudinal research. In a typical study, independent variables causally associated with a specific disease endpoint, such as cholesterol or systolic blood pressure with coronary heart disease, are measured in a population and correlated with the occurrence of the endpoint. Many disease states take years to manifest and are of relative low incidence so longitudinal research must be conducted on large populations and can take 5-10 years to record a significant number of endpoints. Once a significant number of endpoints has been collected a predictive model can be generated. This analysis involves comparing the characteristics of those subjects who

acquired the disease with those who did not.

Classic examples of multivariate predictive equations developed from longitudinal research are the multivariate logistic regression equations for predicting coronary heart disease and stroke from the Framingham Heart Study. The coronary heart disease prediction model from Framingham has come to be regarded as a gold standard for predicting an individual's probability of heart disease onset. (4). However, even the researchers of Framingham admit to limitations of constructing predictive equations from longitudinal research (4, 5). New epidemiological factors have been elucidated since the study was initiated. These new risk factors include serum albumin, plasma fibrinogen, lipoprotein (a), homocysteine, and C-reactive protein (6-8) and were not evaluated in the original cohort of Framingham. Although these factors are currently being measured in the ongoing Framingham population, it will take another 5 to 10 years before existing predictive equations can be updated to include these factors. It is expected, however, that 5 to 10 years from now the epidemiology of heart disease will continue to unfold and the "updated" regression equations will again become obsolete .

Another problem in constructing multivariate regression equations is assigning the predictive weight to each variable. Not all human studies seem to agree on which variables or risk factors demonstrate valid correlations with a disease endpoint and the quantitative extent of each factor with the endpoint. The correlation of homocysteine with the onset of coronary heart disease is a case in point. Its value as a CHD risk factor is being actively debated (9-11) in part because the results from well designed studies has been so equivocal (12-15). Meta-analysis is the

technique of choice for summarizing the strength and consistency of correlation coefficients for a single variable across multiple studies. However, meta-analysis does not provide tools to combine different independent variables from disparate studies. The challenge is to develop an approach for building regression models that does not require 5-10 years of longitudinal data collection and can immediately reflect all current validated medical findings. Synthesis analysis is a new statistical method that can address this challenge.

SYNTHESIS ANALYSIS

Methodology

The process of developing a synthesis analysis equation is a two-stage methodology once a clinical endpoint and a respective array of candidate variables are identified. In the first stage the association of each variable with the clinical endpoint is assessed from published epidemiological research using conventional meta-analysis tools. In the second stage the association of the collective array with the endpoint is determined with the synthesis analysis procedure. This method accounts for the

colinearity among the array of variables. The multivariate equation that predicts the clinical endpoint from a group of factors ultimately becomes the prediction model. Figure 1 illustrates the two stage process.

In the figure, Y represents the outcome variable of interest, in this case the likelihood of disease onset. The four risk factors, X1 to X4, are believed to be associated with the onset of the disease. Suppose that the joint association of X1, X2 and X3 with the outcome Y has been well established. New research findings suggest that X4 is also associated with Y. Synthesis analysis combines the association of X4 with Y into the established association of Y with X1 to X3.

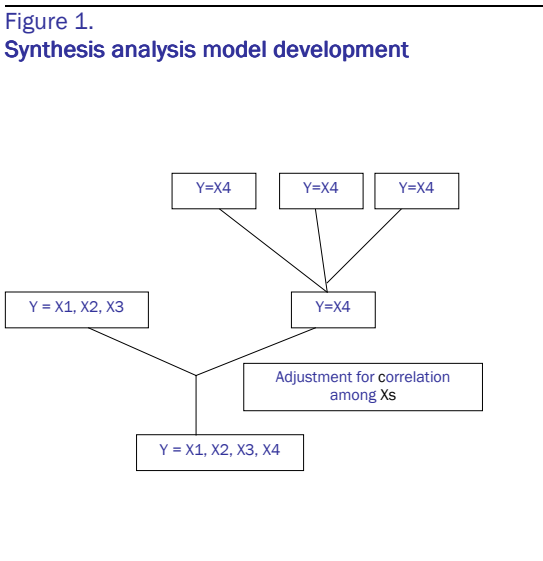
Step one of the process is to summarize the association of Y with X4 through meta-analysis. In this example, the consensus association of Y with X4 is derived from the combination of the three independent studies.

Step two is to integrate the association of X4 with Y into the established association of Y with X1 to X3 through synthesis analysis. During the integration process the correlations among all the Xs derived from a cross-sectional study of a representative population are statistically accounted for.

Underlying assumptions

The synthesis analysis methodology is predicated on three general assumptions: (a) the validity of each relative risk generated in epidemiological studies; (b) the validity of the mathematical principles used in the synthesis analysis process to construct an algorithm; (c) the inferential validity of the population sampling assumptions.

Since synthesis analysis derived predictive equations are constructed from disparate research, the validity of the respective correlation coefficients for each independent variable is critical. In general,



only results from large prospective cohorts are used during the meta-analysis of each independent variable. For example, approximately 40 well-recognized cohorts were used during synthesis analysis to construct an equation for CHD onset. These cohorts included the Framingham Heart Study, the Atherosclerosis Risk in Communities (ARIC) study, the Multiple Risk Factor Intervention Trial (MRFIT), and the NHANES I Epidemiologic Follow-up Study (NHEFS). The first assumption does not assert that the synthesized equation is somehow more valid than the underlying literature; only that it is not less valid.

Validation of the mathematical model is demonstrated by proving the algebraic identity of the algorithm (19). As in all multivariate regression analysis, in synthesis analysis variables are added one at a time into the predictive equation. The basis of the proof is that for any given variable to be added to the equation, its correlation with the previously calculated probability is tantamount to all colinearity with the previously entered variables.

Given the validity of the correlation coefficients and certain mathematical principles used in the synthesis methodology, the cohorts themselves must be representative of both the populations used by the synthesis analysis technique for cross-sectional correlations and the target populations for which the predictive equations are intended. The first assumption is that the individual associations of the Xs with Y and the correlations among the Xs used in the analysis are all representative of the same underlying population. Since each longitudinal study intends to reflect certain biological associations that should be applicable to the general population, this assumption is not difficult to meet. The second more practical assumption is that such representative correlations can be

reasonably determined for a specific population. For this purpose we assume that the Third National Health and Nutritional Examination Survey (NHANESIII) is an adequate dataset in which to determine the correlations among the Xs in the American public and in Western societies in general. The NHANES surveys are nationally representative surveys that are conducted for the express purpose of determining the state of health of all Americans. Considerable effort is made in the design and implementation of the survey to assure that they are broadly representative of the American public.

Current medical understanding of disease risk factors comes from studies that, for the sake of inference, claim to be representative of a larger, more general population. The biomedical community generally agrees with these assertions, otherwise these studies would have been long ago discredited. It is not a difficult next step to assume that an analysis that spans all of these studies, that already claim wide inferential power, is itself in possession of wide inferential power. The statistical assumption that the NHANESIII survey is an adequate database for determining the needed correlations is also not difficult to grasp. The painstaking design of the NHANES series is intended so that health information may be inferred about the entire American public. If one agrees that the Centers for Disease Control and Prevention have done a reasonable job of meeting this goal then this assumption is valid.

SYNTHESIS ANALYSIS VALIDATION

Given the validity of these underlying assumptions, the following exercises were designed to validate synthesis analysis-derived equations for predicting CHD onset.

Predicting CHD onset with a synthesis analysis-derived equation

Using data from the Framingham Heart Study (n=5209), individual probabilities of CHD onset were estimated and compared using, 1) an empirically derived multivariate logistic regression equation and, 2) a synthesis analysis-derived predictive equation. Both equations were constructed using the variables age, sex, smoking status, total cholesterol (chol) and systolic blood pressure (sbp). The empirical logistical regression equation for CHD onset was:

$$\text{logit}(p_1) = -4.6225 + 0.0142 * \text{age} + 0.8595 * \text{sex} - 0.0623 * \text{smoke} + 0.00674 * \text{cholesterol} + 0.0107 * \text{sbp}$$

P1 represents the probability of developing CHD within 5 years.

The synthesis analysis technique was used to build a similar logistic regression equation utilizing the univariate regression coefficients derived from the Framingham Study, together with a correlation matrix among the five variables derived from NHANESIII data

$$\text{logit}(p_2) = -0.9765 + 0.7876 * \text{sex} + 0.0316 * (\text{age} - 44) + 0.005863 * (\text{cholesterol} - 224) + 0.002 * (\text{sbp} - 137) - 0.002 * (\text{smoke} - 0.577)$$

The two equations formulated from the same variables but by different statistical methods were applied to middle-aged NHANESIII subjects (n=8939) and disease onset probabilities were analyzed. The correlation coefficient for the probabilities from the two prediction equations was found to be 0.95. A t-test of the means showed P1 and P2 were not statistically different.

Two conclusions can be drawn from this exercise: (a) the disease onset probabilities calculated by the synthesis analysis derived equation are a reliable representation of the

corresponding disease onset probabilities calculated from the empirically derived equation, and (b) the assumption that the correlations among variables are statistically stable between representative populations is valid.

In a second exercise the number of individuals predicted to develop CHD within 5 years was estimated with a logistic regression equation from the Framingham study (4) and a multivariate logistic equation derived by synthesis analysis. Both equations should predict the same number of CHD events for the target population

The synthesis analysis derived equation was constructed by adding family history of CHD, physical exercise level, serum albumin, plasma fibrinogen, C-reactive protein, homocysteine, aspirin use, hormone replacement therapy, and lipoprotein(a) variables to the Framingham model which included age, gender, smoking status, diabetes, systolic blood pressure, the ratio of total and HDL cholesterol, and left ventricular hypertrophy.

The Framingham and synthesis models were applied to the NHANESIII data. The study subjects were grouped according to different risk factors and the total numbers of subjects predicted to be CHD cases in 5 years (the sum of the probabilities for all the subjects) were compared for the two prediction equations. Table 1 shows that both equations predicted the same number of CHD cases for the target population. More specifically, when the target population was stratified by gender, age, smoking status, or cholesterol ratios, both equations predicted the same number of CHD cases for each subpopulation. Thus, again the synthesis analysis-derived equation provides a reliable representation of disease onset probabilities when compared to the probabilities estimated

from the empirically derived Framingham equation.

Table 1

Number of people predicted to have CHD in five years among NHANESIII population by Framingham CHD model and Synthesis Analysis CHD model.

	Framingham Equation	Synthesis Equation
Gender		
Male	180	183
Female	372	372
Age		
35-44	23	23
45-54	73	72
55-64	139	138
65-74	182	183
75-84	135	138
Smoking		
Yes	387	389
No	166	166
Total:HDL Cholesterol		
Quartile		
1 st	202	201
2 nd	163	164
3 rd	118	119
4 th	69	70

Improved discrimination for CHD onset with synthesis analysis-derived equations

If the synthesis analysis methodology represents a valid approach for constructing predictive equations, and, if the technique is used to update a validated and accepted prediction equation from the Framingham study, then the updated equation would be expected to demonstrate increased CHD onset discriminatory power over the Framingham model.

First, the discriminatory benefit of updating the Framingham equation with a single new risk variable was considered. The relative risk for lipoprotein(a) was added to the published Framingham regression equation using the synthesis analysis procedure (See Table 2). The average probability of CHD onset across quartiles of serum lipoprotein(a) was examined. While the synthesis analysis-derived regression equation produced a gradient of CHD risk across quartiles, the Framingham equation did not detect the direction of either the risk gradient or its magnitude. Thus, the synthesis analysis methodology for adding the variable lipoprotein(a) to the Framingham equation is validated and the discriminatory power of the Framingham equation is strengthened.

Table 2.

The average probability of CHD onset across quartiles of serum lipoprotein(a) using the Framingham or Synthesis Analysis-derived logistic regression equation. The population represents middle-aged subjects of NHANESIII.

Lipoprotein(a) Quartile	Framingham Equation	Synthesis Equation
1 st	7.48	5.04
2 nd	7.19	6.35
3 rd	7.12	8.28
4 th	6.71	10.32

Next, we compared the comprehensive synthesis-analysis derived equation with the Framingham equation in terms of the degree of concordance in CHD risk classification. The synthesis equation includes 8 additional risk variables. We arbitrarily chose a risk of CHD greater than 10% in 5 years as high risk. Both equations were used to predict CHD risk with middle-age NHANESIII subjects and those classified as high risk were identified from each model and then compared between models.

The results are shown in Table 3. While both models agreed on a low risk assignment of 84% of the subjects, there was considerable discordance in which subjects were at high risk when the synthesis model, with the 7 added risk factors, was compared with the Framingham model. If the synthesis-analysis equation truly is more discriminatory than the Framingham equation, then these observations are expected and the synthesis method used to construct the more discriminatory equation is valid.

Table 3.

Comparison of Framingham and Synthesis CHD models in identifying subjects at high risk (>10% risk of CHD in 5 years) among middle-age NHANESIII subjects.

		Synthesis Equation		
		Normal	High Risk	Total
Framingham Equation	Normal	7545	349	7894
	High Risk	284	765	1049
	Total	7829	1114	8943

PREDICTING STATIN DRUG EFFICACY WITH SYNTHESIS ANALYSIS BASED MULTIVARIATE STATISTICS

Pravastatin is approved by the FDA for the primary prevention of myocardial infarction and cardiovascular mortality. Clinical trial data used to support these indications included morbidity and mortality endpoint analysis of drug and placebo populations from the West of Scotland Prevention Study Group (WOSCOPS) (20). We designed a

simulation analysis to investigate the ability of a synthesis analysis-derived predictive equation to assess pravastatin efficacy.

Population demographics of the WOSCOPS cohort were used to select a simulation population from the NHANESIII survey. This population was first used to simulate a clinical trial placebo group by quantitating the population probability of coronary heart disease onset using the comprehensive synthesis analysis-derived predictive equation described in this paper. Next, the same population was used to simulate a clinical trial pravastatin group by lowering the population’s cholesterol values to the same level observed in the WOSCOPS cohort and then reassessing the population risk of CHD onset. Additionally, it has been reported that statins also decrease systolic blood pressure and plasma fibrinogen (21, 22). These parameters were also reduced to reported levels in the simulated pravastatin group.

The predictive equation estimated pravastatin would reduce the risk of CHD by 38% compared to an actual 36% risk reduction observed in the WOSCOPS morbidity and mortality endpoint study. It is interesting to note that when the multivariate prediction was based only on a reduction in cholesterol compared with a reduction in cholesterol, fibrinogen, and systolic blood pressure, a 28% reduction in CHD risk was estimated. These findings are consistent with the theory that surrogate endpoints based on more than one pharmacodynamic property will more closely capture or predict true clinical benefit. These preliminary findings also validate the synthesis analysis methodology for constructing multivariate predictive equations.

Table 4.

Comparison of methods used to assess Pravastatin efficacy in reducing primary risk of coronary heart disease.

Population	Method	CHD Risk Reduction
WOSCOPS	Clinical endpoint	36 %
NHANESIII	Surrogate endpoint (Multivariate Equation)	38 %

MULTIVARIATE PREDICTIVE STATISTICS IN DRUG DEVELOPMENT

The pharmaceutical and biotech industries are constantly striving to shorten the drug development lifecycle and determine early on which drug candidates should move forward in the process. In addition there is increasing demand to demonstrate economic benefit to third party payers. Predictive statistics can play a variety of important roles in shortening research timelines given a practical means for building and applying predictive tools.

Synthesis analysis brings a level of practicality to applying predictive statistics to drug development. With this technique a multivariate predictive equation can be constructed from an array of independent variables that has been defined *a posteriori* to conducting any longitudinal research. Prediction equations are synthesized from completed longitudinal research and can be constructed from disparate studies focused on a common clinical endpoint. Thus, there is now a practical means for developing multivariate prediction equations at relatively low cost and in a timely manner. The need to conduct years of longitudinal research to correlate a specific array of variables with a given endpoint is no longer necessary.

This new technology will allow drug development scientists to build the most

discriminatory multivariate equations for predicting specific clinical endpoints from all of the available longitudinal research and clinical trial studies. Equations can be updated quickly to include new biochemical, clinical, pharmacodynamic, and pharmacogenomic findings, continuously increasing the ability to predict true clinical endpoints.

Multivariate statistics in clinical development: Surrogate endpoints and endpoint modeling

The role of predictive statistics in drug development will be, in part, driven by how accurately statistics can predict true clinical endpoints and the regulatory impact of predictive surrogates in clinical development. Historically, the FDA has relied on morbidity and mortality endpoints for measuring drug safety and efficacy. However, surrogate endpoints have been accepted as an alternative providing sufficient validation studies have been conducted regarding the surrogate's ability to reasonably likely predict clinical benefit (16).

Most surrogate endpoints in use today focus on assessing the efficacy of a drug intervention on a single intermediate in the causal pathway of the true or intended clinical endpoint. For example, statins represent a class of therapeutic agents that are thought to reduce CHD morbidity and mortality by reducing atherosclerotic plaque via a reduction in the causal intermediate serum cholesterol. In this surrogate strategy statins are approved in anticipation of providing the intended clinical benefit of reducing CHD morbidity and mortality, based on the drugs direct effect on the surrogate endpoint cholesterol.

A valid concern regarding surrogate endpoints in clinical development is that the surrogate may not always mimic the drug's efficacy for the intended true clinical

endpoint and surrogates may not address questions of drug safety. Multivariate predictive statistics may represent the next stage in the evolution of surrogate endpoints. In this paper we present preliminary evidence from a drug efficacy modeling simulation that multivariate prediction can accurately assess drug effect on reducing CHD onset. This was based on the analysis of an array of intermediate factors (surrogate array) and other variables causally linked to the onset of CHD. Using the NHANESIII cohort to simulate an experimental population of subjects enrolled in a pravastatin intervention trial, we estimated a 25% reduction in risk of CHD compared to a 36% reduction actual observed in the clinical trial. This estimation only takes into account the drug effect on reducing serum cholesterol, however it has been reported that statins also reduce systolic blood pressure and serum fibrinogen. A 38% reduction in risk of CHD was observed using a multivariate prediction equation when these intermediates are considered.

This preliminary evidence suggests that surrogate endpoints derived from the multivariate analysis of a surrogate array can very closely predict true clinical benefit. In addition when multiple pharmacodynamic properties of a drug are considered in designing the surrogate array, the accuracy of the surrogate endpoint increases. Synthesis analysis statistics is ideally suited to update and improve the accuracy of statistically-based surrogate endpoints as drug pharmacodynamic and pharmacogenomic properties are elucidated.

Given sufficient validation, our results support the notion that surrogate endpoints based on an array of casual intermediates will more closely capture the effect of treatment on the true clinical endpoint compared to endpoints based on a single

assessment. Sopko and Friedman present a conceptually similar strategy with the “surrogate equation” (23)

In addition to its role as a surrogate endpoint, and perhaps of more immediate application, synthesis analysis-based multivariate prediction can provide a variety of support roles in clinical development:

- (1) **Clinical development strategy: modeling drug efficacy.** Modeling the potential drug efficacy for secondary and/or new indications from the analysis of ongoing or retrospective clinical trial data. Supporting go/no go clinical development strategic decisions.
- (2) **Late-stage clinical development: efficacy trend analysis:** Interim analysis of phase III clinical trial data to evaluate efficacy trends.
- (3) **Early-stage clinical development:** Supporting clinical development decisions to move from phase II to phase III based on multivariate surrogate endpoint analysis.

Conceptually, multivariate prediction can be applied to any clinical endpoint or disease state where epidemiological factors, drug pharmacodynamics, and other independent variables in the causal disease process have been characterized. Synthesis analysis provides a practical and heretofore unavailable methodology to define and construct multivariate prediction equations from completed longitudinal and clinical research. This should greatly expand the role of predictive statistics in drug development

Multivariate predictive statistics in pharmaco-economic and clinical outcomes research

Defining the clinical benefit of a given pharmacological intervention can provide

extremely valuable information for a variety of outcomes focused investigations including:

- (1) Assessing the cost-effectiveness or cost-utility of an intervention in Pharmacoeconomic research
- (2) As a basis for making formulary management decisions
- (3) Determining clinical benefit in Phase IV post marketing surveillance studies
- (4) Determining population risk reduction strategies in disease management initiatives.
- (5) Establishing drug use guidelines

However, the time and cost requirements to conduct outcomes research on morbidity and mortality endpoints can be prohibitive. Although a surrogate endpoint may be less prohibitive the surrogate may not always capture the entire effect of the drug intervention on the intended true clinical outcome. Multivariate predictive statistics may represent the next step in the evolution of using surrogate endpoints to conduct outcomes research.

For example entire classes of cardiovascular agents, such as drugs indicated for hypertension or hypercholesterolemia, are actually approved based on surrogate endpoints and the anticipated effect on cardiovascular morbidity and mortality outcomes. We suggest these validated surrogate endpoints can be incorporated with other causal variables into a surrogate array, which we demonstrate in this paper can effectively predict drug effect on morbidity outcomes. Evaluating drug effects using a surrogate array and multivariate statistics can be accomplished with relatively small populations in a few months. Thus, the pharmacoeconomic benefit for reducing

morbidity and mortality outcomes can be investigated with minimal cost and time.

Multivariate statistics will allow the design of numerous short-term studies to investigate the economics of combination therapy on reducing morbidity and mortality outcomes. For example, determining the most cost effective combination of antihypertensive and cholesterol-lowering agent on reducing morbidity outcomes such as coronary heart disease and stroke would be impractical given the number of available therapies and the cost to conduct clinical endpoint studies. In this paper we present a surrogate endpoint strategy, multivariate predictive statistics, that can capture the combined effect of a cholesterol-lowering agent and an antihypertensive on CHD morbidity on a small population in a short timeframe. We have designed other multivariate equations that can predict drug benefit on stroke and type II diabetes outcomes.

This cost effective surrogate tool could greatly assist disease management initiatives which frequent address the comorbidity associated with chronic disease onset. Multivariate prediction can be an extremely cost effective means to investigate the pharmacoeconomic benefit of mono and combination therapy on multiple morbidity and mortality endpoints.

Multivariate statistics in clinical trial population targeting and selection

Defining a homogenous subject population is of primary importance when attempting to minimize the effect of confounding variables on determining drug efficacy. Multivariate statistics could be used to enroll a homogenous population of subjects based on the absolute probability of disease onset. For example, subjects with a uniformly high-risk of CHD may represent a more homogenous population of subjects compared to using serum cholesterol and

risk factor categories to enroll patients in a coronary heart disease prevention trial. Further, morbidity and mortality endpoint trials are dependent on observing a minimal number of clinical events. Multivariate statistics can be used to maximize the enrollment of a population that will experience the intended primary clinical endpoint. For example, our findings show that 10-20% of the American population between the ages of 35 and 70 with borderline high-to-high cholesterol (200-240 mg/dl) are actually at relatively low risk for CHD. Multivariate statistics may represent a strong support tool to select patients truly at risk for disease. This strategy not only could reduce the number of patients required to demonstrate an outcome, but may also shorten the trial duration.

MULTIVARIATE PREDICTIVE STATISTICS IN HUMAN DIAGNOSTICS

Determining an individual's likelihood of developing specific chronic diseases so that individualized intervention plans can be designed to prevent disease onset or delay disease progression is pivotal to the goal of disease prevention and preventive, personalized medicine. The challenge for the medical professional is to analyze patient assessments within the context of any available clinical guidelines, estimate the patient's cumulative level of danger for disease onset, and make a determination of to what degree this danger can be reduced after a successful intervention plan. Although individual patient assessments can be quantitative, estimating a patient's cumulative danger for disease onset and to what degree this danger can be reduced after intervention is largely qualitative.

The absolute likelihood of disease onset based on statistical correlations of individual epidemiological factors with a specific clinical outcome, is a standardized and scientifically accepted means for

summarizing and predicting risk of disease onset. In the paper we describe an innovative statistical approach to synthesizing multivariate predictive equations from completed longitudinal research. Because the technique allows variables from disparate research to be combined, existing multivariate predictive equations can be quickly updated with new epidemiological evidence without the need for initiating a new lengthy longitudinal study. Along these lines we have updated the Framingham Heart Study multivariate equations for predicting CHD and stroke onset with several new epidemiological findings including physical exercise level, serum albumin, plasma fibrinogen, C-reactive protein, homocysteine, aspirin use, hormone replacement therapy, and lipoprotein(a). Our findings suggest that about 7% of the adult American population's CHD risk status is more accurately assessed by the addition of these new epidemiological variables to the Framingham equation. Thus, not only is there a method for the standardized calculation of individual risk scores, this score can be updated to increase its discriminatory power with new medical findings.

The American Heart Association's Task Force on Risk Reduction reviewed the important role of risk scores, such as absolute risk or relative risk of disease onset, in developing clinical plans for risk factor management, patient education, and motivation (18). The challenge is to translate all of the available evidence into a single risk score that can provide a realistic picture of the patient's cumulative danger for disease onset.

Synthesis analysis will play a significant role in human diagnostics by providing a practical means to constantly incorporate medical discoveries into simple predictive analytical tools for the daily practice of

medicine. Synthesis analysis is a simple but elegant statistical method that can be used to create multivariate predictive equations when all of the available evidence regarding disease onset has not been correlated in a single longitudinal study. As such, synthesis analysis is a practical means for applying the value of new disease onset and mortality discoveries to the daily practice of medicine.

The intersection of human diagnostics and human genomics has led to several downstream disciplines namely, diagenomics, pharmacogenomics, and proteomics, all of which need to translate discovery into pragmatic tools for predicting and diagnosing polygenic chronic diseases. Currently, synthesis analysis is the only practical approach to combine the predictive power of new discoveries from the fields of human diagnostics and genomics with previously validated disease variables, into continuously expanding and more accurate predictive systems.

REFERENCES

1. Peterson, K. W., and Hilles, S. B. (eds.). SPM Handbook of Health Risk Appraisals. Charlottesville, VA: Occupational Health Strategies, Inc., 1996.
2. Schoenbach, V. J., Wagner, E. H., and Beery, W. L. Health risk appraisal: review of evidence for effectiveness. *Health Serv. Res.* **22**:553-580 (1987).
3. Goetz, A. A., and McTyre, R. B. Health risk appraisal: some methodologic considerations. *Nurs. Res.* **30**:307-313 (1981).
4. Anderson, K. M., Wilson, P. W. F., Odell, P. M., and Kannel, W. B. An updated coronary risk profile: a statement for health professionals. *Circulation* **83**:356-362 (1991).
5. Wilson, P. W. F. Established risk factors and coronary artery disease: the Framingham Study. *Am. J. Hypertens.* **7**:7S-12S (1994).
6. Kullo, I. J., Gau, G. T., and Tajik, A. J. Novel risk factors for atherosclerosis. *Mayo Clin. Proc.* **75**:369-380 (2000).
7. Harjai, K. J. Potential new cardiovascular risk factors: left ventricular hypertrophy, homocysteine, lipoprotein (a), triglycerides, oxidative stress, and fibrinogen. *Ann. Intern. Med.* **131**:376-386 (1999).
8. Ridker, P. M. Evaluating novel cardiovascular risk factors: Can we better predict heart attacks? *Ann. Intern. Med.* **130**:933-937 (1999).
9. Brattström, L., and Wilcken, D. E. L. Homocysteine and cardiovascular disease: cause or effect? *Am. J. Clin. Nutr.* **72**:315-323 (2000).
10. Ueland, P. M., Refsum, H., Beresford, S. A. A., and Vollset, S. E. The controversy over homocysteine and cardiovascular risk. *Am. J. Clin. Nutr.* **72**:324-332 (2000).
11. Scott, J. M. Editorial: Homocysteine and cardiovascular risk. *Am. J. Clin. Nutr.* **72**:333-334 (2000).
12. Chasan-Taber, L., Selhub, J., Rosenberg, I. H., Malinow, R., Terry, P., Tishler, P. V., Willett, W., Hennekens, C. H., and Stampfer, M. J. A prospective study of folate and vitamin B₆ and risk of myocardial infarction in US physicians. *J. Am. Coll. Nutr.* **15**:136-143 (1996).
13. Evans, R. W., Shaten, B. J., Hempel, J. D., Cutler, J. A., and Kuller, L. H. Homocyst(e)ine and risk of cardiovascular disease in the Multiple Risk Factor Intervention Trial. *Arterioscler. Thromb. Vasc. Biol.* **17**:1947-1953 (1997).
14. Folsom, A. R., Nieto, J., McGovern, P. G., Tsai, M. Y., Malinow, M. R., Eckfeldt, J. H., Hess, D. L., and Davis, C. E. Prospective study of coronary heart disease incidence in relation to fasting total homocysteine, related genetic polymorphisms, and B vitamins: the Atherosclerosis Risk in Communities (ARIC) Study. *Circulation* **98**:204-210 (1998).
15. Bostom, A. G., Silbershatz, H., Rosenberg, I. H., Selhub, J., D'Agostina, R. B., Wolf, P. A., Jacques, P. F., and Wilson, P. W. F. Nonfasting plasma total homocysteine levels and all-cause and cardiovascular disease mortality in elderly Framingham men and women. *Arch. Intern. Med.* **159**:1077-1080 (1999).
16. Downing, G. (ed). Biomarkers and surrogate endpoints: Clinical research and applications. Proceedings of the NIH-FDA conference held on 15-16 April 1999 in Bethesda, Maryland, USA.

17. Fleming, TR, DeMets, DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* **125**:605-613 (1996)
18. Grundy, SM, et al., Primary prevention of coronary heart disease:Guidance from Framingham. A statement for healthcare professionals from the AHA Task Force on Risk Reduction. *Circulation* **97**:1876-1887 (1998)
19. Hu, Guizhou, Root, M., System and method for predicting disease onset. United States Patent 6,110,109. (2000)
20. West of Scotland Prevention Study Group. Influence of pravastatin and plasma lipids on clinical events in the West of Scotland Coronary Prevention Study (WOSCOPS). *Circulation* **97**:1440-1445, 1998
NIH
21. LaRosa JC, He J, and Vupputuri S. Effect of statins on risk of coronary disease: a meta-analysis of randomized controlled trials. *JAMA* **282**:2340-2346, 1999
22. Glorioso N, Troffa C, Filigheddu F, Dettori F, Soro A, Parpaglia PP, Collatina S, and Pahor M. Effect of HMG-CoA reductase inhibitors on blood pressure in patients with essential hypertension and primary hypercholesterolemia. *Hypertension* **34**:1281-1286, 1999

**FINAL DRAFT
PRE-PUBLICATION
COPY**