Combining information from multiple data sources to create multivariable risk models: Illustration and preliminary assessment of a new method

Authors: Samsa G, Hu G, and **Root, M**

Abstract

A common practice of metanalysis is combining the results of numerous studies on the effects of a risk factor on a disease outcome. If several of these composite relative risks are estimated from the medical literature for a specific disease, they cannot be combined in a multivariate risk model, as is often done in individual studies, because methods are not available to overcome the issues of risk factor colinearity and heterogeneity of the different cohorts. We propose a solution to these problems for general linear regression of continuous outcomes using a simple example of combining two independent variables from two sources in estimating a joint outcome. We demonstrate that when explicitly modifying the underlying data characteristics (correlation coefficients, standard deviations, and univariate betas) over a wide range, the predicted outcomes remain reasonable estimates of empirically derived outcomes (gold standard). This method shows the most promise in situations where the primary interest is in generating predicted values as when identifying a high-risk group of individuals. The resulting partial regression coefficients are less robust than the predicted values.

Greg Samsa, Guizhou Hu, and **Martin Root**(2005) Combining Information From Multiple Data Sources to Create Multivariable Risk Models: Illustration and Preliminary Assessment of a New Method. *Journal of Biomedicine and Biotechnology* (ISSN 1110-7243) copy available @ (http://downloads.hindawi.com/journals/jbb/2005/524952.pdf)

Journal of Biomedicine and Biotechnology • 2005:2 (2005) 113–123 • DOI: 10.1155/JBB.2005.113

RESEARCH ARTICLE

Combining Information From Multiple Data Sources to Create Multivariable Risk Models: Illustration and Preliminary Assessment of a New Method

Greg Samsa,[1,2] Guizhou Hu,[3] and Martin Root[3]

[1]*Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, USA*
[2]*Center for Clinical Health Policy Research, Duke University Medical Center, Durham, NC 27705, USA*
[3]*BioSignia, Inc, 1822 East NC Highway 54, Durham, NC 27713, USA*

Received 5 February 2004; revised 29 March 2004; accepted 6 April 2004

A common practice of metanalysis is combining the results of numerous studies on the effects of a risk factor on a disease outcome. If several of these composite relative risks are estimated from the medical literature for a specific disease, they cannot be combined in a multivariate risk model, as is often done in individual studies, because methods are not available to overcome the issues of risk factor colinearity and heterogeneity of the different cohorts. We propose a solution to these problems for general linear regression of continuous outcomes using a simple example of combining two independent variables from two sources in estimating a joint outcome. We demonstrate that when explicitly modifying the underlying data characteristics (correlation coefficients, standard deviations, and univariate betas) over a wide range, the predicted outcomes remain reasonable estimates of empirically derived outcomes (gold standard). This method shows the most promise in situations where the primary interest is in generating predicted values as when identifying a high-risk group of individuals. The resulting partial regression coefficients are less robust than the predicted values.

INTRODUCTION

We propose essentially a multivariate metanalytic technique. Many diseases have numerous risk factors, which are often studied in diverse cohorts with only a limited number of risk factors in each. We here propose a method of combining univariate relative risks (betas) from diverse studies into multivariate models.

Metanalysis has proven to be a powerful tool, when handled appropriately, to summarize previous medical research on a common topic, including epidemiologic research [1, 2, 3]. Several issues need to be carefully considered in reaching conclusions from the metanalysis of epidemiologic studies. Studies are often heterogeneous in their findings [4], which can even be considered a benefit in understanding the source of differences in research findings [5]. Publication bias must also be evaluated in a field in which the decision to publish, the quality and size of the study, and the publishing journal's reputation are strongly interconnected [6].

All authorities agree that the best means of combining effect estimates is by a pooled analysis where the separate study datasets are combined together with possible confounders [3], especially if this pooling is planned prospectively. In general though, effect estimates (β coefficients) are combined from published reports. Univariate betas are combined as in the example of Ernst et al, who considered fibrinogen as a risk factor for cardiovascular disease using univariate and age-adjusted parameters [7]. More commonly, multivariate-adjusted odds ratios and relative risks are used as by Etminan et al on the effects of NSAIDs on Alzheimer's disease onset [8] by Vincent et al on hypoalbuminemia in acute illness [9], or by Danesh et al in summarizing various plasma risk factors and heart disease [10].

Our method of preparing multivariate risk models by metanalysis suggests a comparison with multivariate metanalysis. Unfortunately this term covers several techniques, none of which are similar to ours. In some cases it refers to a metanalysis that considers several similar outcomes with the same risk factor [11, 12, 13]. Another technique, also called metaregression, is essentially a weighted multivariate analysis of all the confounders (and possible sources of heterogeneity) in the summarized studies [11].

Correspondence and reprint requests to Martin Root, BioSignia, Inc, 1822 East NC Highway 54, Durham, NC 27713, USA, E-mail: mroot@biosignia.com

Most true multivariate metanalyses are pooled analyses of multiple studies together with possible confounders. Farrer et al used a pooled analysis to estimate the effect of the interaction between age, sex, and ethnicity on the effect of apolipoprotein *E*4 as a predictor of Alzheimer's disease [14].

There would be clear benefit to a metanalytic technique that could combine univariate risk factors for a disease obtained from different studies. In the metanalysis of Danesh et al, four blood parameters were determined to be significantly correlated with heart disease risk [10]. However, it was impossible to combine those in any meaningful way to determine their joint predictive power or to determine their independence from one another. Oftentimes a researcher simply wants to add a single risk factor to an established multivariate risk model. In a recent example, coronary artery calcium score was combined with the Framingham score for predicting heart disease [15]. Previous studies had showed that both scores were strongly predictive of heart disease but to combine them took a 7-year study with 1461 subjects on whom both scores were collected.

In another example, Gail et al developed a model for the prediction of breast cancer onset [16]. This was subsequently modified by statisticians of the National Surgical Adjuvant Breast and Bowel Project (NSABP) to define eligibility criteria for the Breast Cancer Prevention Trial [17]. They used a new source of data, the Surveillance, Epidemiology, and End Results (SEER) Program for new incidence rates for invasive breast cancer to replace the original incidence rates for total breast cancer. They also used the same new dataset to determine race-specific incidence rates to replace the Whites-only rate from the original model [18]. Their method is reported in an NSABP document [19]. In a final example, the need was simply to modify the Framingham score [20] for heart disease to a new, lower-risk, population. Based on the assumption that the multivariate betas are similar across populations, some authors have suggested changing the equation intercept to reflect the underlying incidence rate of the new population [21, 22]. Others have suggested also modifying the risk factor values themselves by using the prevalence rates of the new population [23, 24]. All of these examples involve combining evidence from different sources into unique multivariate risk models based broadly on the assumption that the correlations among risk factors and between risk factors and the endpoint were not significantly different between data sources or populations.

Matchar et al used a variety of techniques and datasets to develop the Stroke Prevention Policy Model (SPPM) [25]. While they concede difficulties and shortcomings of such an approach, they conclude, for clinical and economic applications, "that despite the difficulties in developing comprehensive models, ..., the benefits of such models exceed the costs of continuing to rely on more conventional methods." The SPPM was then used in demonstrating the economic benefit of a stroke treatment's short-term effect on long-term economic outcomes [26].

The method we are about to introduce can also be used in datasets with a large fraction of missing values. Generally two strategies exist for such situations, either using modeling techniques to extract data from observations with incomplete data or data imputation [27]. Zhao et al have introduced a joint estimating equation that robustly estimates effect size in multivariate models with missing data [28]. Steyerberg et al address the related problem of underpowered small studies [29]. They describe a method to combine results from the medical literature with results from individual patient data and conclude "that prognostic models {from small studies} may benefit substantially from explicit incorporation of literature data."

We have developed a new statistical method to address the question of combining estimates of partial regression parameters across datasets. This method is intended to provide an approximate solution in the circumstances illustrated above. The performance of this method is assessed via simulation.

METHODS

Notation

The continuous outcome variable is denoted by Y, and its predicted value by $\hat{Y}$. We first consider a "gold-standard" dataset including all the predictors of interest. Information from the gold-standard dataset is denoted with an asterisk. In practice this gold-standard dataset will not be available, and the predictors of interest will be distributed across multiple "candidate" datasets. The goal will be to estimate, using information from the candidate datasets, the regression relationship between the risk factors and the outcome that would have been observed if the complete dataset had been available.

Denote the vector of predictors in the gold-standard dataset by

$$X^* = (X_0^*, X_1^*, \ldots, X_Q^*) \tag{1}$$

with the first element $X_0^* = 1$ being included in order to estimate the intercept and the remaining Q predictors being of primary interest. The multivariable regression of Y^* on X^* is

$$\hat{Y}^* = \hat{\beta}^* X^*, \tag{2}$$

where $\hat{Y}^*$ is the predicted value of Y and β^* is estimated in the usual way as

$$B^* = (X^{*'} X^*)^{-1} (X^{*'} Y^*). \tag{3}$$

In other words, the multivariable regression equation observed in the data is

$$\hat{Y}^* = a^* + b_1^* X_1^* + b_2^* X_2^* + \cdots + b_Q^* X_Q^*, \tag{4}$$

where, for example, b_1^* estimates β_1^*, and so forth. We will focus on the regression coefficients observed in the data—that is, on B^* rather than β^*.

In the above equations, the estimates of the partial regression coefficients produced from the gold-standard dataset are

$$B^* = (b_1^*, b_2^*, \ldots, b_Q^*). \quad (5)$$

In contrast, each of the "univariable" regression coefficients, denoted, for example, by b_{u1}^*, is the result of fitting a univariable regression model with a single predictor—for example,

$$\hat{Y}^* = a_{1u}^* + b_{1u}^* X_1^*. \quad (6)$$

We assume that there are Q univariable regression coefficients available for use, one from each candidate dataset. The vector of univariable regression coefficients from the candidate datasets is denoted as

$$B_u = (b_{u1}, b_{u2}, \ldots, b_{uQ}). \quad (7)$$

In practice, the observed values of B_u and B_u^* can differ because of (a) differences between β_u and β_u^* and (b) sampling variability within each of the datasets in question.

For concreteness, our goal is to estimate the multivariable regression model, as summarized through the set of Q partial regression coefficients $(b_1^*, b_2^*, \ldots, b_Q^*)$ and the predicted values $\hat{Y}^*$ that could have been produced were the gold-standard dataset is available. In the absence of this gold-standard dataset, we assume that from Q candidate datasets, each containing exactly one predictor variable, we have available its standard deviation (for study j, denoted by s_j and combined into a vector S), as well as its univariable regression coefficient (for study j, denoted by b_{uj}, and combined into a vector B_u). We also assume that from one or more additional datasets, the various first-order correlations between each set of predictors (denoted by r_{ij}, and combined into a matrix R) are available. (These additional datasets need not contain Y.)

It is important to note that this formulation of the problem includes, as a special case, the situation where the various studies include overlapping risk factors. In particular, for each study (whose number need not equal Q) we could estimate a set of univariable regression coefficients—that is, one coefficient per risk factor per study. The additional problem induced by overlapping predictors is that different estimates of b_{uj} will be available for some or all of the risk factors, and that each of these estimates must somehow be reconciled into a single "best" estimate. In this case, we might (1) use standard metanalytic techniques to combine the various estimates of b_{uj} or (2) select the b_{uj} from the "best" available datasets. Estimates of S and R that reconcile multiple estimates can be generated in a similar fashion.

Proposed approach

To illustrate our proposed approach, termed the univariable synthesis method, first consider the gold-standard dataset. Within this dataset, we can calculate (1) univariable regression coefficients for each predictor, denoted by B_u^*, (2) standard deviations for each predictor, denoted by S^*, and (3) the set of all pairwise correlations between the predictor variables, denoted by R^*. Denoting element-wise multiplication by "$\cdot$," and element-wise division by "/," the core of the univariable synthesis method relies on noting that $(b_1^*, b_2^*, \ldots, b_Q^*)$—that is, the portion of B^* excluding the intercept—can also be estimated by [30, equation 1]

$$B^* = \frac{(R^{*-1}(B_u^* \cdot S^*))}{S^*}. \quad (8)$$

The basic idea behind the univariable synthesis method is that, when candidate datasets must be used, the various elements of B_u, R, and S can nevertheless be accumulated across these multiple data sources. In order to do so, it must be assumed that the relevant standard deviations, univariable regression coefficients, and correlations are comparable across studies. (More precisely, we are assuming, in analogy to the random-effect model used in metanalysis, that each of the above terms represents a realization from the same superpopulation. Thus, the assumption is not that the various studies are "identical," but rather that they are "similar.") The fundamental insight is that B_u, R, and S are more likely to be similar across datasets than are the partial regression coefficients.

In order to obtain appropriately calibrated values of $\hat{Y}$, an estimate of the intercept of the above multivariable regression model is also required. This can be obtained by forcing the predicted regression function to pass through the point (X_m, Y_m), where X_m is the vector of mean values of the predictors, and Y_m is the mean response.

Assessment

The fundamental assumption of the univariable synthesis method is that the various first-order summary measures B_u, R, and S are comparable across datasets. More precisely, this fundamental assumption holds that the values of B_u, R, and S, obtained from various candidate datasets, are similar to those values of B_u^*, R^*, and S^* that would have been obtained from the gold-standard dataset, if these data were available. If the above inputs are comparable, then applying (8) for B^{*1Q} to the set of first-order summary measures from the candidate datasets is conceptually equivalent to calculating B^{*1Q} from the gold-standard dataset, and thus to recreating the best possible estimate of the desired gold-standard regression model.

The validity of this basic assumption, and thus of the methodology as a whole, can potentially be assessed in two ways. First, we could ask the *empirical* question, namely, *to what degree do estimates of R, B_u, and S tend to be similar across multiple datasets?* (The question of whether B_u is similar across datasets is a standard problem in metanalysis—the question of whether R and S are similar has been less exhaustively studied.) Second, we

could ask the *mathematical* question, namely, *what is the impact, on the partial regression coefficients and predicted values for individual subjects, of discrepancies between the gold-standard estimates of R^*, B_u^*, and S^* and estimates of R, B_u, and S obtained from the candidate datasets?* In other words, we could perform a mathematical sensitivity analysis to determine the degree to which the above discrepancies in the inputs are likely to affect the outputs.

Both assessment approaches suffer from a fundamental difficulty; namely, that the number of potential regression models to which the proposed technique could be applied is infinite (eg, regression models can differ in the number of predictor variables as well as the values of B_u, R, S, and B). Therefore, (a) demonstrating that the method works well in one circumstance does not necessarily demonstrate that it will work well in others; and (b) the number of possible circumstances is so large that it is difficult to develop a set of scenarios that would be sufficiently representative. We deal with this difficulty by setting up a single scenario (described in detail later) that is both simple and typical. Given this scenario, we then perform a mathematical sensitivity analysis across a wide range of parameter values and observe the effects of these changes on (a) the estimated regression coefficients and (b) set of the predictions generated by the model. Though not intended to be a definitive analysis, this approach does allow us to assess the robustness of the methodology in its most basic form; and also to illustrate how the users of this methodology can set up a sensitivity analysis that is tailored to the characteristics of their own data.

Sensitivity analysis methods

The dataset for the sensitivity analysis has 84 subjects and 3 variables: an outcome Y, a commonly accepted predictor X_1, and a new predictor X_2. (The raw data happened to be taken from a study in exercise physiology, but the source is not as important as the fact that X_1 and X_2 operate in exactly the same fashion as risk factors in epidemiologic investigations.) Table 1 provides a list of the data.

The gold-standard multivariable regression, having $R^2 = 0.67$, is

$$\hat{Y}^* = 1743.94 - 92.65X_1^* + 39.44X_2^*. \quad (9)$$

The standard deviations of b_1^* and b_2^* above are 7.93 and 12.07, respectively. The univariable regressions are

$$\hat{Y}^* = 1751 - 76.53X_1^*, \quad (10)$$

where $R^2 = 0.62$ and

$$\hat{Y}^* = 576.19 - 48.34X_2^*, \quad (11)$$

where $R^2 = 0.11$. The standard deviations of these univariable regression coefficients b_{u1}^* and b_{u2}^* are 6.56 and 15.38, respectively. All of the regression coefficients are statistically significant. The correlation between the predictors is 0.62, and the standard deviations of the predictors are 0.94 and 0.62, respectively. In this dataset (a) the commonly accepted risk factor is a relatively good predictor of the outcome; (b) once the commonly accepted risk factor is included in the model, the new predictor has an

Table 1. Raw data used in simulation examples.

y	x_1	x_2	y	x_1	x_2
223.1	19.8	8.3	149.0	21.1	8.6
105.4	21.0	8.5	171.0	20.3	8.8
161.9	21.4	8.8	111.0	21.4	8.9
161.3	21.3	9.0	99.0	21.8	9.2
94.1	21.0	8.2	267.0	19.0	8.4
280.5	19.7	8.3	98.0	21.0	8.6
183.6	19.7	8.0	184.1	19.0	8.4
204.4	21.0	8.7	416.1	19.0	8.4
140.2	20.2	8.1	112.3	20.5	8.6
73.0	21.4	9.1	583.6	19.0	8.5
194.0	20.4	8.4	53.4	21.8	8.9
118.0	21.1	8.6	180.4	19.8	8.1
68.3	22.0	9.6	128.0	21.6	8.8
131.0	21.4	9.1	82.4	21.7	9.1
127.0	21.0	8.5	230.8	20.4	8.7
72.2	21.6	8.9	135.2	21.6	8.9
93.0	21.8	9.5	90.8	22.0	9.6
94.9	21.0	8.9	181.0	20.5	8.8
108.3	22.0	9.2	99.0	21.7	8.7
118.9	20.3	6.7	321.6	19.0	8.6
83.8	20.9	6.6	134.7	21.5	8.7
66.6	22.0	9.9	342.0	19.9	8.4
117.7	21.1	8.6	115.0	20.9	8.6
209.6	19.0	6.3	185.0	20.9	8.7
137.0	20.8	8.5	164.0	20.0	8.9
66.0	21.2	8.8	89.6	22.0	9.8
174.8	21.1	8.4	225.7	20.5	8.5
427.8	19.0	9.0	179.1	20.7	8.5
179.6	21.4	8.9	54.9	22.0	9.3
237.3	19.5	8.0	96.3	21.5	8.3
209.9	19.8	8.4	71.0	21.2	8.9
319.0	19.3	8.1	62.5	22.0	10.0
89.7	21.6	8.6	191.8	19.5	8.1
122.0	22.0	9.4	65.0	21.9	9.2
112.1	22.0	9.2	201.0	21.2	8.8
131.8	21.5	8.7	116.0	21.0	8.7
80.0	22.0	9.6	191.0	20.3	8.3
87.0	21.5	8.6	136.7	21.1	8.8
247.0	19.0	8.3	137.4	21.1	8.8
70.0	21.1	9.2	67.0	22.0	10.0
63.5	22.0	9.3	207.0	19.4	8.4
224.7	20.4	8.7	122.0	21.3	9.3

incremental benefit which is of moderate magnitude; (c) the commonly accepted and new risk factors are positively correlated; and (d) when comparing the multivariable and univariable models, some of the parameter values differ (indeed, the regression coefficient for X_2^* changes sign). These characteristics are present in many epidemiological datasets.

To implement the sensitivity analysis, we modified three of the inputs: (a) the values of R^* were varied by adding from −0.10 to +0.10, in increments of 0.01, to the baseline value of 0.62; (b) the values of B_u^* were varied by adding from −15 to +15, in increments of 1.5, to the baseline values of −76.54 and −48.34; and (c) the values of S^* were varied by adding from −0.15 to +0.15, in increments of .015, to the baseline values of 0.94 and 0.62. The differences between these inputs and the true values from the gold-standard dataset play the role of the variability likely to be observed by using the candidate datasets rather than the gold-standard dataset. (The above perturbations of the inputs were derived on intuitive grounds in order to represent from small to moderately large differences between the above datasets—for example, the extreme values for B_u^* are in the range of 1–2 standard deviations from the values in the gold-standard dataset. In practice, the user might base the choice of perturbations on more substantive considerations pertinent to the scientific issues at hand.)

For each set of simulation inputs, we reestimated the multivariable regression model using (8), thus obtaining the following: (a) new multivariable regression coefficients and (b) new predicted values. To determine how close the new multivariable regression coefficients were to the gold-standard values, we calculated a standardized distance (D) [30]:

$$D = \left\{ \frac{\{[(b_1 - b_1^*)/s(b_1)]^2 + [(b_2 - b_2^*)/s(b_2)]^2\}}{2} \right\}^{1/2}. \tag{12}$$

For example, for the simulation with B_u unchanged, S unchanged, and R increased from 0.62 to 0.66, the estimated partial regression coefficients become −98.87 and 51.36. The standardized distance is

$$\begin{aligned} D &= \left\{ \frac{1}{2} \left(\left[\frac{6.22}{6.56} \right]^2 + \left[\frac{11.92}{15.38} \right]^2 \right) \right\}^{1/2} \\ &= (0.75)^{1/2} = 0.87 \end{aligned} \tag{13}$$

implying that the average change in the partial regression coefficients is a bit less than one standard deviation. To determine how consistent the predicted values were, we took the correlation between $\hat{Y}$ and $\hat{Y}^*$, where $\hat{Y}$ and $\hat{Y}^*$ are the vectors (ie, across all subjects) of predicted outcomes for the two models in question. For the above example, the correlation was 0.997.

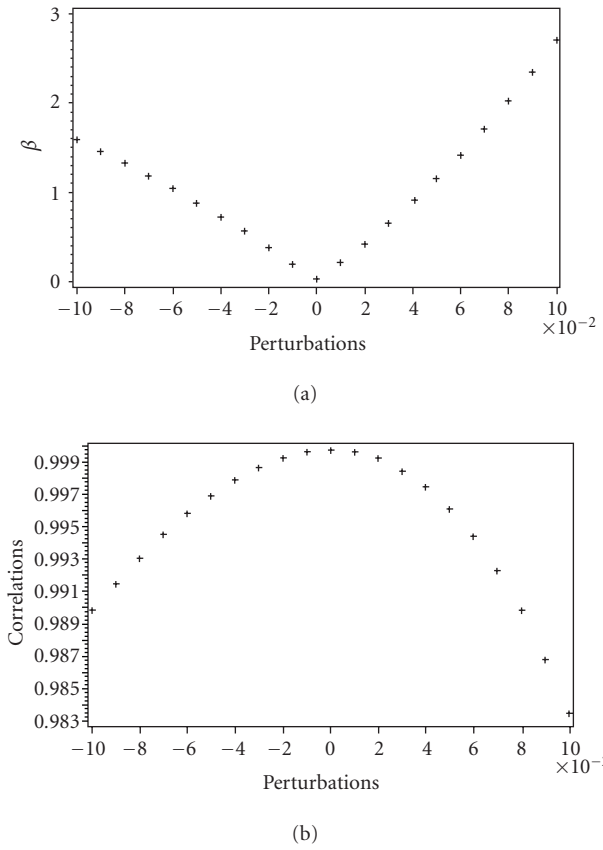

Figure 1. (a) Univariable synthesis method—effect of perturbing R on partial regression coefficients. The x-axis represents the perturbation; the y-axis represents the change in the regression coefficient in standardized distance between the perturbed and unperturbed models. (b) Univariable synthesis method—effect of perturbing R on correlations. The x-axis represents the perturbation; the y-axis represents the correlation between the predicted values for the perturbed and unperturbed models.

RESULTS

Figures 1–5 summarize the results. In particular, each set of figures describes the impact, on either the standardized difference between B and B^* (Figures 1a, 2a, and 3a) or the correlation between $\hat{Y}$ and $\hat{Y}^*$ (Figures 1b, 2b, and 3b), of perturbing one of the inputs, while keeping all other inputs at the true values from the gold-standard dataset. Figure 1 shows the effects of perturbing R. Figure 2 shows the effects of perturbing b_{u2}. Similar results were found for perturbing b_{u1}. Figure 3 shows the effects of perturbing s_1. Similar results were found for perturbing s_2.

Figures 4 and 5 show the effects of perturbing both b_{u1} and b_{u2} on the estimated values of $Y(\hat{Y})$ and on the model residuals compared to the unperturbed model. Similar results were found for perturbing both s_1 and s_2. The residuals from the perturbed models had similar distributions

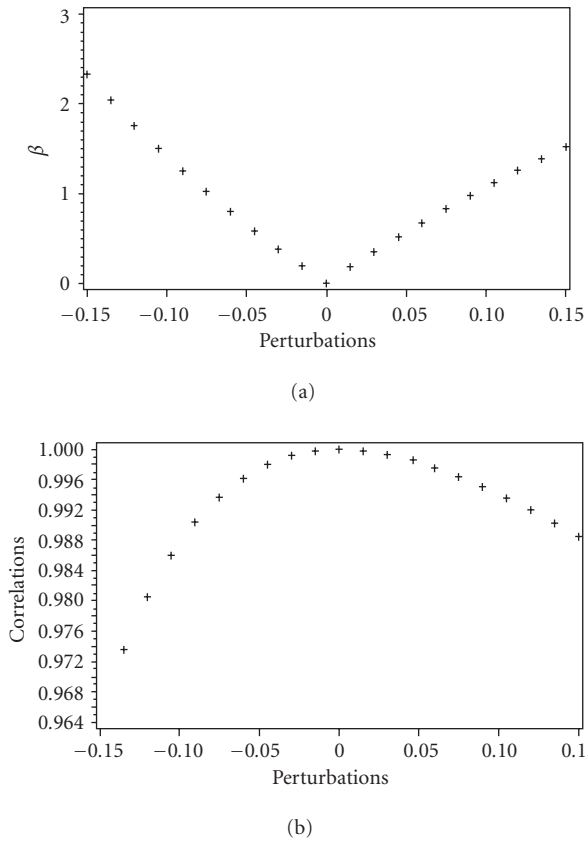

FIGURE 2. (a) Univariable synthesis method—effect of perturbing b_{u2} on partial regression coefficients. The x-axis represents the perturbation; the y-axis represents the change in the regression coefficient in standardized distance between the perturbed and unperturbed models. (b) Univariable synthesis method—effect of perturbing b_{u2} on correlations. The x-axis represents the perturbation; the y-axis represents the correlation between the predicted values for the perturbed and unperturbed models.

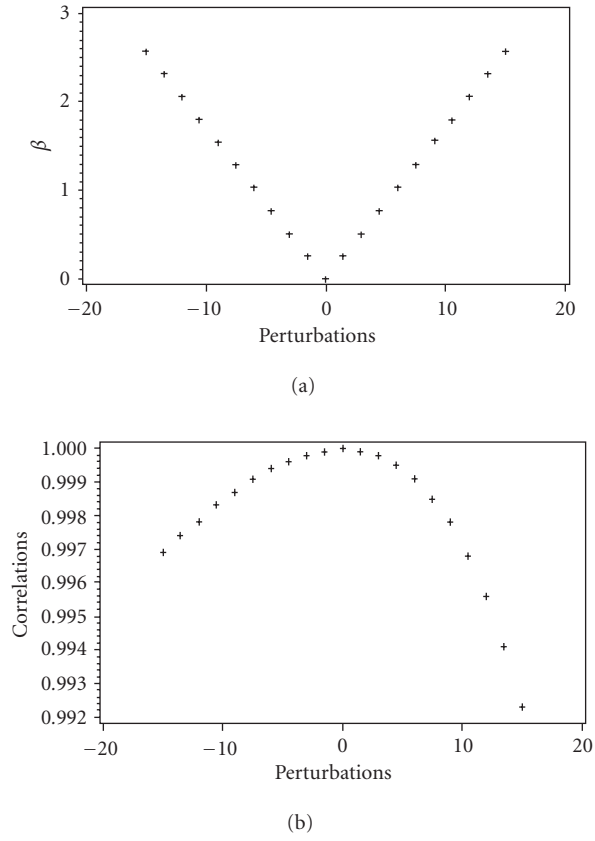

FIGURE 3. (a) Univariable synthesis method—effect of perturbing s_1 on partial regression coefficients. The x-axis represents the perturbation; the y-axis represents the change in the regression coefficient in standardized distance between the perturbed and unperturbed models. (b) Univariable synthesis method—effect of perturbing s_1 on correlations. The x-axis represents the perturbation; the y-axis represents the correlation between the predicted values for the perturbed and unperturbed models.

compared to those of the unperturbed model (data not shown). Also, plots of the residuals from the perturbed and unperturbed models against X_1, X_2, and $\hat{Y}^*$ were very similar (data not shown).

Even modest perturbations of the inputs affect the estimated values of the partial regression coefficients; for example, varying R by 0.05 units is associated with an approximately 1-unit difference between B and B^*. Perturbing the inputs has much less impact on the correlation between the predicted values. For example, applying the above perturbation to R resulted in a correlation between $\hat{Y}$ and $\hat{Y}^*$ exceeding 0.99. Similar results were observed when perturbing all the inputs simultaneously (data not shown).

In summary, the univariable synthesis approach appears to be robust to changes in its inputs, so long as what the user is ultimately interested in is the predicted values resulting from the multivariate regression. The methodology is relatively less robust when estimating the values of the partial regression coefficients.

DISCUSSION

Creating multivariable regression models containing partial regression coefficients is central to the practice of epidemiology. It is quite common for the risk factors (predictors) of interest to be distributed across multiple datasets. Because the value of partial regression coefficients depends upon the choice of the other variables that are included in the model, simply combining partial regression coefficients across datasets may be dangerous. Indeed, combining partial regression coefficients across datasets is the most dangerous in the situation of most practical interest, that is, when the correlations among the risk factors in question are moderate to strong. One strength of the univariable synthesis method is that the

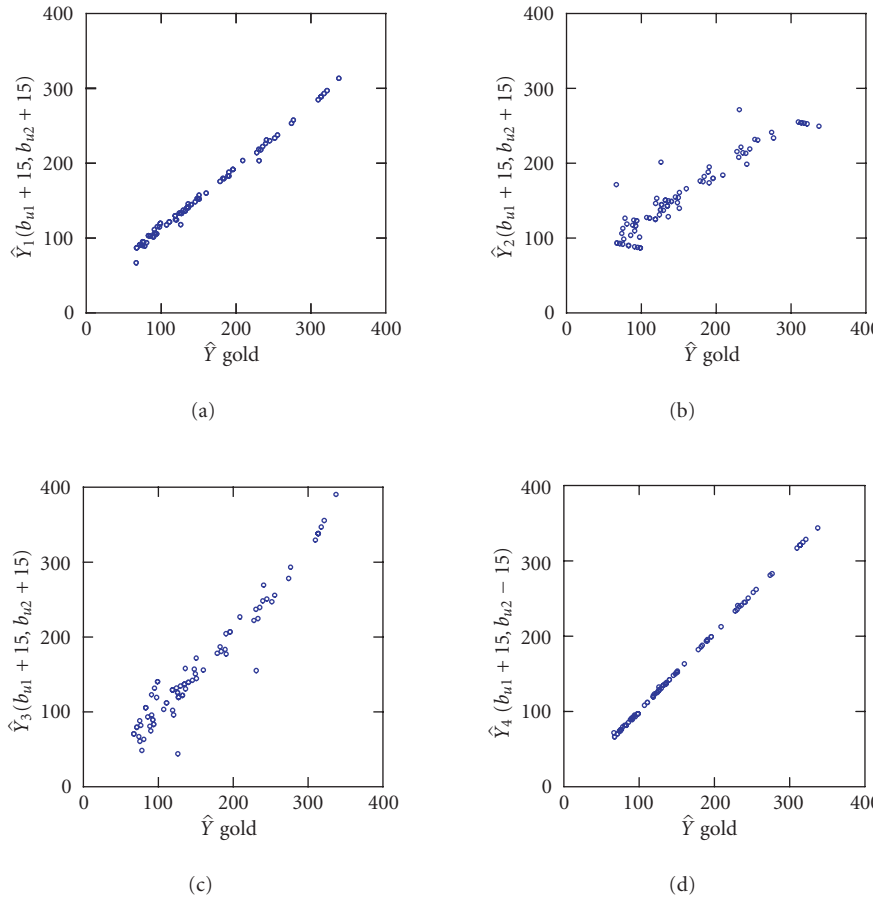

FIGURE 4. Univariable synthesis method—effect of perturbing both b_{u1} and b_{u2} on correlations between the predicted values for the perturbed and unperturbed models. The x-axes represent the estimated Y of the unperturbed model ($\hat{Y}$ gold); the y-axes represent the estimated Y of the perturbed models. The y-axes labels indicate the perturbation. For example, for Y_1, both b_{u1} and b_{u2} were perturbed by adding 15. The estimating equation was then computed and $\hat{Y}_1$ was calculated.

correlations among the predictors are explicitly considered in the quantitative estimation of the partial regression coefficients.

We know of no ideal solution to this problem, but have proposed the univariable synthesis method as a possible way forward. The most critical assumption underlying this method is that first- and second-order information such as univariable regression coefficients, standard deviations, and correlations are comparable across datasets. Admittedly, the assumption of comparability is strong, but it is not essentially different from what must be assumed in order to make qualitative conclusions about epidemiological phenomena based on information from multiple sources, or what must be assumed when information about individual risk factors is quantitatively combined across studies using metanalysis. In any event, it might be argued that (a) these assumptions are being made explicitly rather than implicitly; (b) sensitivity analyses can be performed in order to assess the impact of these assumptions; and (c) the alternatives—namely, ignoring the issue entirely or limiting the number of risk factors to be modeled—have significant difficulties of their own.

The univariable synthesis method has a number of limitations. As discussed above, it assumes that first- and second-order information can be combined across datasets. Fortunately, the technique appears to be reasonably robust to modest departures from these assumptions—particularly when the focus of inference is on the predictions generated by the model rather than the parameter estimates themselves. Other limitations include the inability to deal with interactions and the difficulty of generating estimates of precision (eg, standard errors of regression coefficients).

A limitation of our assessment is the less-than-comprehensive nature of the sensitivity analyses. In essence, by selecting a single dataset to use as an archetype, we implicitly assume that the goal of the sensitivity analyses is demonstration of the plausibility of the concept rather than definitive proof. This is a generic problem in

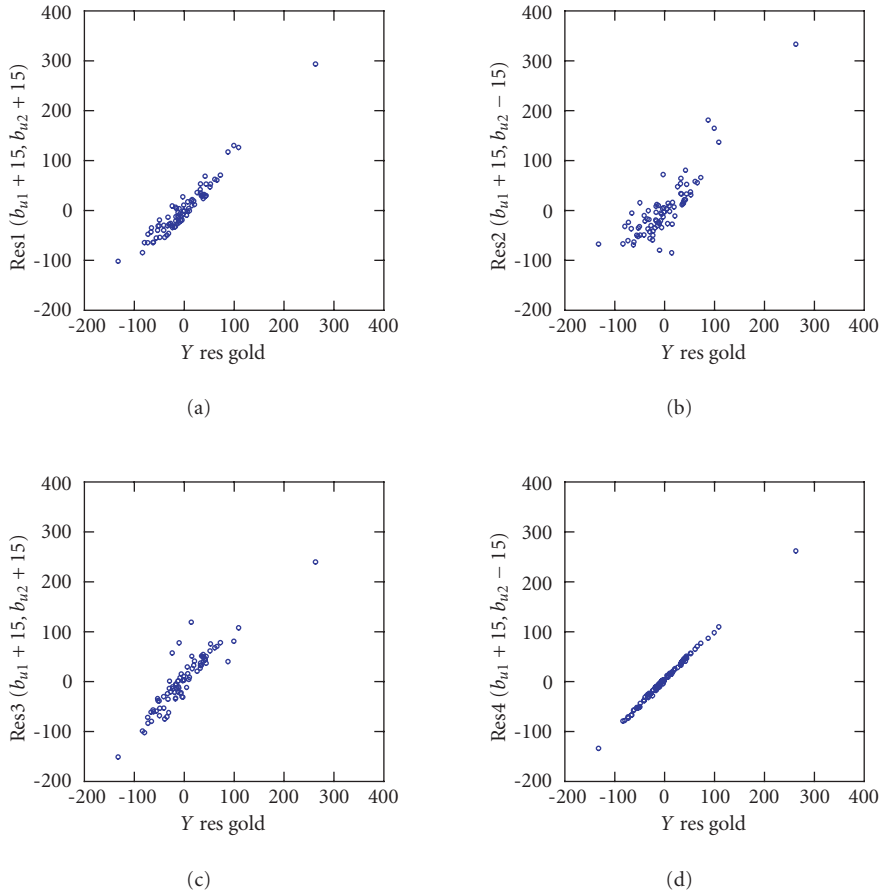

FIGURE 5. Univariable synthesis method—effect of perturbing both b_{u1} and b_{u2} on residuals of the perturbed and unperturbed models. The x-axes represent the residuals of the unperturbed model (Y res gold); the y-axes represent the residuals of the perturbed models. The y-axes labels indicate the perturbation. For example, for RES1, both b_{u1} and b_{u2} were perturbed by adding 15. The estimating equation was then computed and Y_1 residual was calculated.

the use of simulation methodology to analyze the properties of statistical methods having application across a wide range of conditions.

An implication of the above is that before using the univariable synthesis method in practice, the user should always perform a sensitivity analysis relevant to his or her application. The observed data should be assumed to represent the gold-standard, and the implications of permuting the inputs to the synthesis analysis techniques can be assessed as illustrated here. (Thus, a further assumption is being made—namely, that the local behavior of the system near the values of the gold-standard estimates can be adequately modeled by the local behavior of the system near the sampled values from the candidate datasets.)

A final limitation applies to those applications where the regression coefficients are of more interest than the predicted values. The univariable synthesis method is more robust with respect to its predicted values than to the values of its regression coefficients. In large part, this may simply be a reflection of the general instability of partial regression coefficients.

Under what circumstances might the univariable synthesis method be applied? Perhaps the most natural application would be to generate lists of patients at high-risk. For example, a predicted length of stay for post-stroke rehabilitation could be generated, the 10% of patients with the highest predicted lengths of stay could be identified, then be targeted for an intervention intended to reduce this length of stay. Such an application focuses much more on predicted values than regression coefficients, and thus makes use of the component of this methodology with the greatest apparent robustness. In this case, the interpretation of the simulation results indicates that the correlation between the gold-standard and the candidate datasets becomes critical. For example, (assuming a normal distribution of predicted values) if this correlation is 0.95, 0.97, and 0.99, then, of those patients with the highest 10% of predicted values generated by the univariable synthesis methodology, approximately 79%, 83%, and 91% of patients will be in the top 10% generated from the gold-standard database. (If the distribution of predicted values has heavier tails than the normal distribution, then

these percentages will be even higher.) Thus, the magnitude of correlations observed in our simulations implies that those patients identified by the univariable synthesis method as having extreme predicted values of the outcome are likely to be actually extreme.

One principle that is implicit in the above discussion is that, because the univariable synthesis method is assumption intensive, in any given circumstance its application will involve a trade-off between its approximate nature (a negative) and the improvement in prediction obtained by being able to include additional risk factors (a positive). Thus, this trade-off would be most likely to favor the adoption of the new method in situations where (a) substantive considerations suggest that the various candidate datasets are comparable (ie, thus reducing the negative impact of the assumptions); (b) the new predictors explain a substantively important amount of the variation outcome, above and beyond the traditional predictors (ie, thus, increasing the positive impact of being able to include new predictors); and (c) the primary focus is on the predicted values themselves rather than the models' partial regression coefficients (ie, because the robustness of the method is greatest for their predicted values). Encouragingly, these conditions describe a significant area of epidemiological practice, especially if the set of potential outcomes is expanded to include dichotomous outcomes (such as the incidence of disease) and time until survival. Extensions of the univariable synthesis and related methods to other types of outcomes will be presented elsewhere. Ongoing challenges for the developers involve both extending these methods and determining the set of applications for which these new tools are best suited.

APPENDICES

A SAS implementation of univariable simulation method

The input file, named inputs, contains x_0, x_1, x_2, and y;

```
proc iml;
  use inputs (keep=y);
  read all  into y;

  use inputs (keep=x_0);
  read all into x0;

  use inputs (keep=x_1);
  read all into x1;

  use inputs (keep=x_2);
  read all into x2;

  x1x2=x1||x2;
  x0x1=x0||x1;
  x0x2=x0||x2;
  x=x0||x1||x2.
```

This portion of the code generates the standard regression results;

```
  xpxi=inv(t(x)*
  x);
  beta=xpxi*(t(x)*y);
  yhat=x*beta;
  resid=y-yhat;
  sse=ssq(resid);
  n=nrow(x);
  dfe=nrow(x)-ncol(x);
  mse=sse/dfe;
  cssy=ssq(y-sum(y)/n);
  rsquare=(cssy-sse)/cssy;
  r=corr(x1x2);
  stdb=sqrt(vecdiag(xpxi)*mse);

  beta1=inv(t(x0x1)*x0x1)*(t(x0x1)*y);
  beta2=inv(t(x0x2)*x0x2)*(t(x0x2)*y);
```

This portion of the code implements the univariable synthesis approach;

```
  s1=sqrt((ssq(x1-sum(x1)/n))/(n-1));
  s2=sqrt((ssq(x2-sum(x2)/n))/(n-1));

  bu=beta1[2,1] // beta2[2,1];
  s=s1 // s2;
  invr=inv(r);
  bus=bu#s;
  invrbus=invr*bus.
```

b_syn is the estimated regression coefficient, yhat_syn is the predicted outcome, where yhat_syn can be further modified to lie on the line with slope b_syn and passing through the point consisting of the means of all variables;

```
  b_syn=(inv(r)*(bu#s))/s;
  yhat_syn=x1x2*b_syn;


print b_syn yhat_syn;

quit;
 run.
```

B Illustration of the calculations for the univariable synthesis method

From the dataset in Table 1,

$$R = \begin{bmatrix} 1.0000000 & 0.6223841 \\ 0.6223841 & 1.0000000 \end{bmatrix},$$

$$B_u = \begin{bmatrix} -76.52528 \\ -48.34035 \end{bmatrix},$$

$$S = \begin{bmatrix} 0.9403281 \\ 0.6177001 \end{bmatrix}.$$

The steps in the calculation are as follows:

$$R^{-1} = \begin{bmatrix} 1.6322853 & -1.015908 \\ -1.015908 & 1.6322853 \end{bmatrix},$$

$$B_u \cdot S = \begin{bmatrix} -71.95886 \\ -29.85984 \end{bmatrix},$$

$$R^{-1}(B_u \cdot S) = \begin{bmatrix} -87.12253 \\ 24.363844 \end{bmatrix},$$

$$(R^{-1}(B_u \cdot S))/S = \begin{bmatrix} -92.65121 \\ 39.442839 \end{bmatrix}.$$

ACKNOWLEDGMENT

This research was funded by BioSignia Inc, which placed no limitations on publication.

REFERENCES

[1] Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev.* 1987;9:1–30.

[2] Egger M, Schneider M, Smith GD. Spurious precision? Meta-analysis of observational studies. *BMJ.* 1998;316(7125):140–144.

[3] Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol.* 1999;28(1):1–9.

[4] Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol.* 1994;140(3):290–296.

[5] Mosteller F, Colditz GA. Understanding research synthesis (meta-analysis). *Ann. Rev. Public Health.* 1996;17:1–23.

[6] Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *BMJ.* 2001;323(7304):101–105.

[7] Ernst E, Resch KL. Fibrinogen as a cardiovascular risk factor: a meta-analysis and review of the literature. *Ann Intern Med.* 1993;118(12):956–963.

[8] Etminan M, Gill S, Samii A. Effect of non-steroidal anti-inflammatory drugs on risk of Alzheimer's disease: systematic review and meta-analysis of observational studies. *BMJ.* 2003;327(7407):128–132.

[9] Vincent JL, Dubois MJ, Navickis RJ, Wilkes MM. Hypoalbuminemia in acute illness: is there a rationale for intervention? A meta-analysis of cohort studies and controlled trials. *Ann Surg.* 2003;237(3):319–334.

[10] Danesh J, Collins R, Appleby P, Peto R. Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies. *JAMA.* 1998;279(18):1477–1482.

[11] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med.* 2002;21(4):589–624.

[12] Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Stat Med.* 2003;22(14):2309–2333.

[13] Arends LR, Vokó Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. *Stat Med.* 2003;22(8):1335–1353.

[14] Farrer LA, Cupples LA, Haines JL, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA.* 1997;278(16):1349–1356.

[15] Greenland P, LaBree L, Azen SP, Doherty TM, Detrano RC. Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals. *JAMA.* 2004;291(2):210–215.

[16] Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst.* 1989;81(24):1879–1886.

[17] Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst.* 1999;91(18):1541–1548.

[18] Fisher B, Costantino JP, Wickerham DL, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst.* 1998;90(18):1371–1388.

[19] Anderson S, Ahnn S, Duff K. NSABP Breast Cancer Prevention Trial Risk Assessment Program (Version 2). NSABP Biostatistical. *Center Technical Report.* 1992.

[20] Anderson KM, Wilson PW, Odell PM. An updated coronary risk profile. A statement for health professionals. *Circulation.* 1991;83(1):356–362.

[21] Laurier D, Nguyen PC, Cazelles B, Segond P., Estimation of CHD risk in a French working population using a modified Framingham model. The PCV-METRA Group. *J Clin Epidemiol.* 1994;47(12):1353–1364.

[22] Menotti A, Lanti M, Puddu PE, Kromhout D. Coronary heart disease incidence in northern and southern European populations: a reanalysis of the seven countries study for a European coronary risk chart. *Heart.* 2000;84(3):238–244.

[23] D'Agostino R, Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA.* 2001;286(2):180–187.

[24] Marrugat J, D'Agostino R, Sullivan L. Adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. *J Epidemiol Community Health.* 2003;57(8):634–638.

[25] Matchar DB, Samsa GP, Matthews JR, et al. The Stroke Prevention Policy Model: linking evidence and clinical decisions. *Ann Intern Med.* 1997;127:704–711.
[26] Samsa GP, Reutter RA, Parmigiani G, et al. Performing cost-effectiveness analysis by integrating randomized trial data with a comprehensive decision model: application to treatment of acute ischemic stroke. *J Clin Epidemiol.* 1999;52(3):259–271.
[27] Pigott TD. Missing predictors in models of effect size. *Eval Health Prof.* 2001;24(3):277–307.
[28] Zhao LP, Lipsitz S, Lew D. Regression analysis with missing covariate data using estimating equation. *Biometrics.* 1996;52(4):1165–1182.
[29] Steyerberg EW, Eijkemans MJ, van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Stat Med.* 2000;19(2):141–160.
[30] Draper N, Smith H. *Applied Regression Analysis.* 2nd ed. New York, NY: John Wiley & Sons. 1981.