American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Research Article

Copyright© Gefei Li

Development of Composite Index for Multiple Correlated Risk Factors in Clinical Research

Gefei Li* and Shein Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina

***Corresponding author:** Gefei Li, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, 2424 Erwin Road, Durham, North Carolina.

To Cite This article: *Gefei Li* and Shein Chung Chow, Development of Composite Index for Multiple Correlated Risk Factors in Clinical Research. Am J Biomed Sci & Res. 2026 29(5) AJBSR.MS.ID.003838,* **DOI:** 10.34297/AJBSR.2026.29.003838

Received: December 23, 2025; **Published:** January 09, 2026

Abstract

In clinical research, a medical predictive modelling is often performed based on a set of risk factors (predictors) not only to inform disease status but also to predict the performance of clinical outcome for an effective disease management. Under a well-established and validated medical predictive model [1] developed a composite index of two highly correlated predictors regardless they may be positively or negatively and/or linearly or nonlinearly correlated to the clinical outcome or response. In this article, we extend [1] results to multiple correlated predictors in two ways. One is to fully utilize all predictors for development of so-called therapeutic index. The other one is to first

(i) divide all predictors into two groups (e.g., efficacy and safety),

(ii) obtain composite index of respective groups, i.e., efficacy index and safety index, and then

(iii) based on the individual composite index to develop a composite index for benefit-risk assessment [2]. The proposed extended composite indices are evaluated both theoretically and via a clinical trial simulation.

Keywords: Disease Management, Multiple Correlated Predictors, Multiplicative Model, Therapeutic Index, Benefit-Risk Assessment

Introduction

In clinical research, a medical predictive model is often developed using an appropriate statistical model based on some risk factors (predictors) which may be correlated positively or negatively in a linear or nonlinear fashion. In practice, a well-established and validated medical predictive model cannot only be used to inform disease status but also provide valuable information regarding disease management including prevention, diagnosis, and treatment of the disease under study. *Li and Chow* suggested building a medical predictive model with a multivariate set of predictors using a (logistic) regression analysis approach by the following steps:

(i) identifying potential predictors (e.g., demographics or patient characteristics) by determining associations between the potential predictors and the response,

(ii) testing for co-linearity among the identified predictors,

(iii) conducting predictive model fitting with the identified predictors,

(iv) performing goodness-of-fit of the fitted model, and

(v) validating the developed medical predictive model both internally (i.e., reproducibility) and externally (i.e., generalizability).

Chow, et al., [1] indicated that most commonly used composite indices in clinical research are of the form of $x_1^a x_2^b$, where x_1 and x_2 are identified highly correlated risk factors (predictors). For example, Body Mass Index (BMI) is commonly considered in obesity studies, where BMI is defined as the ratio between body weight (kg) and the square of height (m), i.e., BMI$= x_1^a x_2^b$, where x_1 is body weight (kg) and x_2 is height (m) with a = 1 and b = -2. For another example, consider studies for examination of QT interval prolongation for cardiotoxicity. The commonly considered index is Bazett 's QT interval adjusted for heart rate (RR), denoted by $QT_C B$, where $QT_C B = x_1^a x_2^b$, where x_1 is QT interval and x_2 is RR with a = 1 and b = -1/2 [3]. Along this time, *Chow, et al.,* [1] proposed the development of a composite index of two highly correlated risk factors under a multiplicative model. *Chow, et al.,* [1] also indicated that their proposed composite index has the following advantages: first, in the interest of parsimony of predictors, the development a composite index reduces a multiple-parameter (e.g., two predictors as discussed in this article) problem to a single parameter (the developed composite index) problem. Second, the developed composite index is able to address the positively/negatively and/or linearly/non-linearly correlation between each of the two predictors (which are correlated each other) and the response. Third, the developed composite index outperforms each individual predictor in two ways:

(i) if each predictor can inform the disease status or treatment effect, the composite index can definitely do and

(ii) if the composite index can inform the disease status or treatment effect, each individual predictor may not be able to.

The purpose of this article is to develop a statistically principled framework for constructing composite indices from multiple, potentially correlated risk factors, and to investigate how such indices can be used for clinical prediction, therapeutic evaluation, and benefit–risk assessment. Specifically, we aim to

(i) propose a systematic procedure for deriving the exponent parameters of a composite index through a log–linear modeling approach, and

(ii) explore practical considerations such as interpretability, rounding of exponent parameters, and implementation in clinical decision making. The remainder of this article is organized as follows. Section 2 briefly outlines the general statistical methodology for developing a composite index based on multiple correlated risk factors, including model formulation, parameter estimation, and the Bayesian predictive framework. Section 3 validates the developed composite index and examines its characteristics and practical challenges. Section 4 presents potential applications, with emphasis on (i) the development of a therapeutic index and (ii) a composite index for benefit–risk assessment. Section 5 reports simulation results and discusses adjacent-integer considerations.

Statistical Method

Notations

Under the multiple regression framework, the dataset can be represented in matrix form as follows:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix},$$

where Y is the n x 1 vector of observed responses, and X is the $n \times k$ design matrix of predictors. The element x_{ij} denotes the value of the i -th predictor for the j -th observation. Without loss of generality, assume that Y and all predictors X_i (i = 1, 2, ..., k) are standardized variables. Under standardized variables, the mean and variance of the regressors are given by

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} = 0, \quad s_i^2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ij} - \bar{X}_i\right)^2 = 1, \quad for\, i = 1,\ldots,k.$$

Similarly, for the standardized variable of the clinical outcome response (dependent variable), we have

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} y_j = 0, \quad s_y^2 = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \bar{Y}\right)^2 = 1.$$

Suppose we are interested in developing a composite index for a group of m highly correlated variables, where m ≤ k.

Under standardized variables, the sample covariance between any two predictors X_i and X_t $(1 \le i, t \le k)$ is given by

$$s_{it} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ij} - \bar{X}_i\right)\left(x_{tj} - \bar{X}_t\right) = \frac{1}{n}\sum_{j=1}^{n} x_{ij} x_{tj}.$$

Similarly, the sample covariance between Y and X_i is

$$s_{iy} = \frac{1}{n}\sum_{j=1}^{n} x_{ij} y_j.$$

As a result, the sample correlation between X_i and X_t and between X_i and Y can be expressed as

$$r_{it} = \frac{s_{it}}{\sqrt{s_i^2 s_t^2}} = s_{it}, \quad r_{iy} = \frac{s_{iy}}{\sqrt{s_i^2 s_y^2}} = s_{iy}.$$

Statistical Model

Under the framework of standardized variables, statistical model can be written as follows

$$Y = X\beta + \varepsilon,$$

where Y is the $n\times 1$ vector of dependent variables, X is the $n\times K$ matrix of regressors, β is the $K\times 1$ vector of regression coefficients, and ε is the $n\times 1$ vector of random errors. Thus,

the Ordinary Least Squares (OLS) estimator of β is given by

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y.$$

Based on standardized variables, $\hat{\beta}$ can be expressed as a function of the sample correlations.

Denote by x_j' the j-th row vector of X. Then the (i,t)-th element of $X'X$ is given by

$$\left(X'X\right)_{it} = \sum_{j=1}^{n} x_{ij}x_{tj} = ns_{it} = nr_{it}.$$

Furthermore, the i-th element of $X'Y$ is

$$\left(X'Y\right)_{i} = \sum_{j=1}^{n} x_{ij}y_{j} = ns_{iy} = nr_{iy}.$$

Now, let r_{xx} denote the sample correlation matrix of X, whose (i,t)-th entry is r_{it}. Thus, r_{xx} is a $k\times k$ symmetric matrix. Similarly, let r_{yX} denote the k×1 vector whose i-th element is r_{iy}. Then we can write

$$X'X = nr_{xx}, \quad X'Y = nr_{yX}.$$

Substituting into the OLS estimator, we obtain

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y = \left(nr_{xx}\right)^{-1}\left(nr_{yX}\right) = r_{xx}^{-1}r_{yX}.$$

Hence, under standardized variables, the OLS estimator depends only on the correlation structure among predictors and their correlations with the response variable. The vector

$$\hat{\hat{a}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k\right)'$$

provides the estimated contribution of each predictor to the response Y, which will be used to identify highly correlated subsets of predictors for composite index construction.

Development of Composite Index

Chow, et al., [1] proposed a general methodology for development of a composite index of two dependent predictors regardless they are positively or negatively and/or linearly or nonlinearly correlated under a well-established and validated medical predictive model. In this subsection, we will focus on the development of a composite index for multiple predictors (i.e., k > 2).

Let $Y \in R^n, X = [X_1, \ldots, X_k] \in R^{n\times k}\left(standardized\ columns\right), \beta \in R^k.$. Fit the single global OLS once:

$$\hat{a}_{all} = \left(X'X\right)^{-1} X'Y\left(equivalently, with\ standardized\ data, \hat{a}_{all} = R_{xx}^{-1}r_{yX}\right).$$

This yields $\hat{\beta}_j$ for every $j=1,\ldots,k.$

Following similar idea as *Chow, et al.,* [1], the composite index of k multiple predictors can be obtained by following the following steps.

Step 1. Identify Highly Correlated Predictors

We first identify subsets of predictors that exhibit high degrees of intercorrelation, as these are the best candidates for integration into a composite index. Specifically, among the k available predictors, we consider all possible subsets of size m = 2,3, 4 … k. The number of such subsets is given by the binomial coefficient $\binom{m}{k}$. For each subset of m predictors, we calculate its correlation coefficient matrix, denoted by A. For this purpose, two complementary methods are proposed to evaluate the degree of collinearity within each subset:

Eigenvalue-based method (Belsley, 1980) – By definition, the correlation between x_i and x_j, denoted by A can be expressed as

$$A = Corr\left(x_i, x_j\right) = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1m} \\ r_{21} & 1 & r_{23} & \cdots & r_{2m} \\ r_{31} & r_{32} & 1 & \cdots & r_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{m1} & \cdots & \cdots & \cdots & 1 \end{pmatrix}$$

For each correlation matrix A, we compute its eigenvalues λ_1, λ_2 ... λ_m

$$|A - \ddot{e}I| = 0 \Rightarrow \ddot{e} = \left(\ddot{e}_1, \ddot{e}_2, \ldots, \ddot{e}_m\right),$$

The condition index κ_{max} is defined as:

$$\hat{e}_{max} = \sqrt{\ddot{e}_{max} / \ddot{e}_{min}}$$

According to Belsley (1980), if $\kappa_{max} < 10$, collinearity can be considered negligible.

If $10 \leq \kappa_{max} \leq 30$, moderate collinearity exists; and if $\kappa_{max} > 30$, the subset exhibits severe collinearity. We calculate κ_{max} for all $\binom{m}{k}$ possible subsets.

To operationalize the eigenvalue-based identification of highly collinear predictors for composite index construction, we consider two complementary subset selection rules:

Thresholded Lexicographic Condition-Index Rule (Diagnostic Rule): We evaluate the condition index $\kappa_{max}(S)$ for all non-empty subsets S of the K candidate predictors. Following the diagnostic interpretation of the condition index [4], we restrict attention to subsets exhibiting moderate multicollinearity, defined as those with $\kappa_{max}(S) > K_0$, We set $K_0 = 10$ as a conservative diagnostic cutoff, restricting attention to subsets that already exhibit at least moderate collinearity; this ensures the selected subset is sufficiently redundant to be well summarized by a single composite index. Among all such subsets, we select the subset with the smallest cardinality |S| thereby identifying the minimal group of predictors that already displays collinearity. If multiple subsets share this minimal size, we retain the subset with the largest condition index $\kappa_{max}(S)$.

Penalized Condition-Index Rule (Complexity–Collinearity Trade-Off) – When no subset satisfies the severe-collinearity threshold, or when one wishes to explicitly balance collinearity strength against subset size, we adopt a penalized selection criterion. Specifically, for each non-empty subset S, we compute $\kappa_{max}(S)$ and define the penalized objective

$$J_\lambda(S) = log(\kappa_{max}(S) ,$$

where $\lambda \geq 0$ controls the penalty on subset size. We then select the subset

$$S^*(\lambda) = arg\,max\, J_\lambda(S).$$

This criterion Favors subsets that yield a substantial increase in the condition index while discouraging mechanically larger subsets. The logarithmic transformation stabilizes the scale of $\kappa_{max}(S)$ and yields an interpretable trade-off: adding one additional predictor must increase $\kappa_{max}(S)$ by at least a factor of e^λ to offset the penalty. In empirical applications, λ can be selected via external validation (e.g., maximizing predictive reproducibility).

Determinant-Based Method – Alternatively, for each correlation matrix A, we compute its determinant, denoted by |A|. The determinant provides a scalar measure of multicollinearity:

$|A| \to 0$ implies strong collinearity,

whereas $|A| \to 1$ implies weak collinearity.

When comparing subsets of different cardinalities, we further consider a dimension-normalized determinant defined as $|A|^{1/m}$, which removes the mechanical dependence of the determinant on subset size and enables meaningful comparison across subsets of different dimensions. In practice, we evaluate this criterion over all non-empty subsets of the k candidate predictors. Specifically, for each subset size $m = 1, \ldots\ , k,$, we enumerate all $\binom{m}{k}$ possible subsets and compute the corresponding correlation matrices and their determinants. The final subset used for composite index construction is selected by minimizing the dimension-normalized determinant $|A|^{1/m}$ across all candidate subsets.

Step 2. Retain the Corresponding Coefficients from the Global Fit

For the chosen m variables, let $S \subset \{1, \ldots, k\}$ denote the index set of the m variables selected in Step 1; in what follows we work with the submatrix X_S and retain the corresponding coefficients $\hat{\beta}_S = \{\hat{\beta}_j : j \in S\}$ from the global fit.

Define the subset-driven linear predictor (partial contribution):

$$\hat{Y}_m = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \hat{\beta}_{m_1} \begin{pmatrix} x_{m_1,1} \\ x_{m_1,2} \\ \vdots \\ x_{m_1,n} \end{pmatrix} + \hat{\beta}_{m_2} \begin{pmatrix} x_{m_2,1} \\ x_{m_2,2} \\ \vdots \\ x_{m_2,n} \end{pmatrix} + \cdots + \hat{\beta}_{m_m} \begin{pmatrix} x_{m_m,1} \\ x_{m_m,2} \\ \vdots \\ x_{m_m,n} \end{pmatrix}$$

$$Let\ X_{m_1} = \begin{pmatrix} x_{m_1,1} \\ x_{m_1,2} \\ \vdots \\ x_{m_1,n} \end{pmatrix} \ldots X_{m_m} = \begin{pmatrix} x_{m_m,1} \\ x_{m_m,2} \\ \vdots \\ x_{m_m,n} \end{pmatrix}$$

$$\Rightarrow \hat{Y}_m = \hat{\beta}_{m_1} X_{m_1} + \hat{\beta}_{m_2} X_{m_2} + \cdots + \hat{\beta}_{m_m} X_{m_m}$$

Step 3. Construct the Multivariate Composite Index for m Variables

Consider

$$\hat{Y}_m = X_{m_1}^{a_1} X_{m_2}^{a_2} \cdots X_{m_m}^{a_m} \varepsilon ,$$

After taking log-transformation, we have

$$\log \hat{Y}_m = a_1 \log X_{m_1} + a_2 \log X_{m_2} + \cdots + a_m \log X_{m_m} + \log \varepsilon$$

Let

$$\hat{Y}'_m = \log \hat{Y}_m, \quad X'_{m_1} = \log X_{m_1}, \quad e = \log \varepsilon$$

$$\hat{Y}'_m = a_1 X'_{m_1} + a_2 X'_{m_2} + \cdots + a_m X'_{m_m} + e$$

Estimate the exponents a_j by the method of ordinary least square (OLS), the resultant composite index is then given by

$$Composite\ Index = \hat{Y}_m = X_{m_1}^{\hat{a}_1} X_{m_2}^{\hat{a}_2} \cdots X_{m_m}^{\hat{a}_m}$$

Validation of the Developed Composite Index

We may validate the developed composite index by considering how close an observed response y, its predicted value $\hat{y}_m$ and the predicted value $\hat{y}_m$ (obtained from the fitted regression model of the composite index) are to one another. Specifically, let the predicted value from the regression model be

$$\hat{Y}_m = \hat{\beta}_{m_1} X_{m_1} + \hat{\beta}_{m_2} X_{m_2} + \cdots + \hat{\beta}_{m_m} X_{m_m}$$

and let the predicted value from the composite index be

$$\hat{Y}_m = X_{m_1}^{\hat{a}_1} X_{m_2}^{\hat{a}_2} \cdots X_{m_m}^{\hat{a}_m}$$

To assess the closeness between $\hat{Y}_m$ and $\hat{Y}_m$ we propose two criteria based on either the absolute difference or the relative difference between them.

$$Criterion\, I.\, p_1 = P\left(\left| \frac{\hat{y}_m - \hat{y}_m}{\hat{y}_m} \right| < \delta \right),$$

$$Criterion\, II.\, p_2 = P\left(\left| \frac{\hat{y}_m - \hat{y}_m}{\hat{y}_m} \right| < \delta \right)$$

where δ denotes a clinically or scientifically meaningful tolerance level. In other words, it is desirable to have a high probability that the absolute or relative difference between the predicted value $\hat{y}_m$ and that from the composite index model $\hat{y}_m$ is less than δ.

Let p_1 and p_2 denote the probabilities defined above. For either $i = 1 or 2$, it is of interest to test the following hypotheses:

$$H_0 : p_i \le p_0 \quad versus \quad H_a : p_i > p_0,$$

where p_0 is a pre-specified constant (e.g., a desired level of predictive agreement).

If the conclusion is to reject H_0 in favour of H_a then the developed multivariate composite index is considered statistically validated.

Internal Validation (Reproducibility Probability)

Consider the following exponential model

$$\hat{Y}_m = X_{m_1}^{a_1} X_{m_2}^{a_2} \cdots X_{m_m}^{a_m} \varepsilon$$

Thus, the constructed composite index with the estimated exponent vector is given by

$$\hat{Y}_m = X_{m_1}^{\hat{a}_1} X_{m_2}^{\hat{a}_2} \cdots X_{m_m}^{\hat{a}_m},$$

where $\hat{a} = \left(\hat{a}_1, \ldots, \hat{a}_m\right)$ is the vector of estimated exponent coefficients obtained from the following log-transformed regression analysis

$$\log \hat{Y}_m = a_1 \log X_{m_1} + a_2 \log X_{m_2} + \cdots + a_m \log X_{m_m} + \log \varepsilon$$

where $Y' = \log\left(\hat{Y}_m\right)$, $X' = \left(\log X_{m_1}, \ldots, \log X_{m_m}\right)$, and $e = \log\left(\varepsilon\right)$.

Thus, we have

$$Y' = X'a + e, \qquad \text{where } e \sim N\left(0, \sigma^2 I\right)$$

The training dataset is given by

$$D = \{\left(X', Y'\right)\}$$

Thus, the reproducibility probability quantifies the chance that the predicted composite index is close to the true composite index for a new patient

$$p_r = P\left(\left| \hat{Y}^* - \hat{Y}^* \right| < \delta | D \right)$$

Substituting the definitions

$$p_r = P\left(\left| \exp\left(Y'^*\right) - \exp\left(X'^{*T} \hat{a}\right) \right| < \delta | D \right),$$

where δ is pre-determined, clinically negligible value.

By Bayes Therom, we have

$$\pi(aD,\sigma)=\frac{p(Y'X',a,\sigma)\pi(a\sigma)}{\int p(Y'X',a,\sigma)\pi(a\sigma)da}$$

where $Y' \mid X',a,\sigma \sim N(X'a,\sigma^2 I)$. If the prior is specified as

$$a \mid \sigma \sim N(\mu_0, \Sigma_0),$$

then the posterior is

$$a \mid D,\sigma \sim N(\mu_a, \Sigma_a)$$

where:

$$\Sigma_a=\left(\frac{1}{\sigma^2}X'^T X'+\Sigma_0^{-1}\right)^{-1},$$

$$\mu_a=\Sigma_a\left(\frac{1}{\sigma^2}X'^T Y'+\Sigma_0^{-1}\mu_0\right)$$

Then the estimate of the exponent coefficients is given by $\hat{a}=\mu_a$.

For a non-informative prior then:

$$\Sigma_a=\left(\frac{1}{\sigma^2}X'^T X'\right)^{-1},\ \mu_a=\frac{1}{\sigma^2}\Sigma_a X'^T Y'$$

For a new patient, the latent log-composite-index value is generated as

$$Y'^*=X'^{*T}a+e^*,\ e^*\sim N(0,\sigma^2)$$

Integrating over the posterior of a yields the Bayesian posterior predictive

$$Y'^* \mid X'^*,D,\sigma \sim N\left(X'^{*T}\mu_a, X'^{*T}\Sigma_a X'^*+\sigma^2\right)$$

The predicted composite index for the new patient is

$$\tilde{Y}^*=\exp\left(X'^{*T}\hat{a}\right)$$

This value is deterministic given data and the new patient's predictors.

The true composite index for the new patient follows the true model:

$$\tilde{Y}^*=X_1^{*a_1}X_2^{*a_2}\cdots X_m^{*a_m}\varepsilon^*,$$

$$Y'^*=\log\left(\tilde{Y}^*\right)=X'^{*T}a+e^*$$

So, the true composite index is:

$$\tilde{Y}^*=\exp\left(Y'^*\right)$$

This is a random variable, whose distribution is given by the posterior predictive above. Then we can compute the reproducibility probability.

For a more realistic case in practice where σ is unknown, we place the following conjugate prior on the log-linear model:

$$a \mid \sigma^2 \sim N(\mu_0,\sigma^2\Sigma_0),\qquad \sigma^2\sim Gamma^{-1}(\alpha_0,\lambda_0)$$

Given the conjugate prior, the posterior distribution is expressed as

$$a \mid X',Y',\sigma^2 \sim N(\mu_a,\sigma^2\Sigma_a),$$

$$\sigma^2 \mid X',Y' \sim Gamma^{-1}(\alpha_\delta,\lambda_\delta),$$

where

$$\Sigma_a=\left(\Sigma_0^{-1}+X'^T X'\right)^{-1},\ \mu_a=\Sigma_a\left(\Sigma_0^{-1}\mu_0+X'^T Y'\right)$$

$$\alpha_\sigma=\alpha_0+\frac{n}{2},$$

$$\lambda_\delta=\lambda_0+\frac{1}{2}\left(Y'^T Y'+\mu_0^T\Sigma_0^{-1}\mu_0-\mu_a^T\Sigma_a^{-1}\mu_a\right)$$

The posterior predictive distribution is given by

$$p\left(Y'^* X'^*,D\right)=\int p\left(Y'^* X'^*,a,\sigma^2\right)p\left(a,\sigma^2 D\right)da\,d\sigma^2$$

$$Y'^* \mid X'^*,D \sim t_{2\alpha_\delta}\left(X'^{*T}\mu_a,\frac{\lambda_\delta}{\alpha_\delta}\left(1+X'^{*T}\Sigma_a X'^*\right)\right)$$

Then we can compute the reproducibility probability.

External Validation (Generalizability Probability)

The posterior distribution a | D δ characterizes how the exponent parameters are distributed in the original population. However, when applying the composite index to a different patient population, the underlying risk structure may change. To evaluate whether the composite index generalizes to such new populations, we adjust the posterior distribution of a to reflect potential population-level differences.

We assume that when moving from the original population to a new population, the exponent parameters may change in two systematic ways:

Mean shift – The average effect of each predictor may differ across populations. This is modeled by a shift vector

$$\lambda=(\lambda_1,\ldots,\lambda_m)^T$$

Variance – Predictor effects may become more or less variable in a new population. To capture this, we introduce a diagonal scaling matrix

$$C=\begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_m \end{pmatrix}, \qquad c_j>0$$

so that the covariance is inflated or deflated by CΣ_a C. Under this population-shift assumption:

$$a^{(new)}\text{쬘}D\,\acute{\Theta}N\left(\mu_a+\lambda, C\ \ {}_aC\right)$$

This distribution represents what the exponent vector would look like if the composite-index model were developed in the new population instead of the original population.

For a new patient with log predictors

$$X^{'*}=\left(\log X^{*}_{m_1},\ldots,\log X^{*}_{m_m}\right)^T$$

the log-composite index in the shifted population becomes:

$$Y^{'*}_{new}=X^{'*T}a^{(new)}+e^{*}, \qquad e^{*}\sim N\left(0,\sigma^2\right)$$

Thus, the posterior predictive distribution is:

$$Y\,\overset{*}{\text{Ọ}}_{new}\text{쬘}X\,{}^{'*},{}^{'}D\sim N\left(\ X\,{}^{'*T}\left(\mu_a+\lambda\right),\ X\,{}^{'\ *T}\left(C\ \ {}_a\,C\right)X\ {}^{*}+\sigma^2\ \right)$$

If σ2 is unknown, the predictive distribution becomes a Student–t distribution, fully analogous to the reproducibility analysis.

True composite index for the new population：

$$\overset{\square}{Y}{}^{*}_{new}=\exp\left(Y^{'*}_{new}\right)$$

Predicted composite index from our model：

$$\overset{\square}{Y}*=\exp\left(X^{'*T}\,\hat{a}\right)$$

Then the generalizability probability is:

$$p_g=P\left(\left|\overset{\square}{Y}{}^{*}_{new}-\overset{\square}{Y}{}^{*}\right|<\delta\,|\,D\right)$$

It measures how likely the composite index, when applied to a different patient population, continues to produce predictions sufficiently close to the true composite-index values of that population.

Remarks

Characteristic of Composite Index

The composite index is data-adaptive, exploiting correlation structure to summarize shared information among predictors. Its multiplicative, log-linear formulation ensures scale invariance and yields interpretable contribution weights. By combining penalized subset selection with external reproducibility evaluation, the index balances parsimony and robustness. Importantly, the composite index retains the original variables and produces directionally meaningful contributions, facilitating clinical interpretation. Unlike Principal Component Analysis (PCA), which generates orthogonal components that are often difficult to interpret clinically, the proposed index preserves the original clinical measurements. In contrast to penalized regression methods such as LASSO, which are primarily designed for sparse prediction, the composite index focuses on stable aggregation of correlated information. Compared with simple unweighted summation, it allows data-driven weighting without imposing equal contributions across predictors.

From a practical perspective, the composite index is straightforward to implement once the relevant predictors are identified. Under a well-established medical predictive model, it offers several advantages, including parsimony in risk factors, potential use as a diagnostic or monitoring tool, and utility in disease management through integration of multiple clinically relevant measurements into a single summary measure [5].

Challenging of Composite Index

Despite its flexibility and interpretability, the proposed composite index faces several practical and methodological challenges. First, its construction relies on the presence of sufficient correlation among candidate predictors. When predictors are weakly related or capture largely independent dimensions, correlation-based screening may fail to identify meaningful subsets, and the resulting index may offer limited advantage over simpler aggregation strategies. In such cases, alternative formulations, such as therapeutic indices based on all available predictors, may be more appropriate. Second, the estimation of exponent coefficients is sensitive to data quality and sample size. As the index is derived from log-transformed variables and fitted through regression-based procedures, small sample sizes, measurement error, or extreme values can disproportionately influence the estimated weights, potentially affecting stability and reproducibility.

Third, predictors may be of different data types, including continuous, categorical, or ordinal variables. In such settings, the proposed methodology may require modification, for example through appropriate encoding or transformation into standardized scores before index construction. verall, these challenges highlight the importance of using composite indices judiciously, with careful attention to data structure, estimation stability, sample size considerations, and application context [6].

Potiential Application

Development of Therapeutic Index

[7] The composite index construction described above relies on

identifying subsets of predictors that exhibit strong intercorrelation. In practice, clinical predictor sets often contain correlated variables, but the therapeutic index retains all available predictors regardless of their dependence structure. A Therapeutic Index (TI) is designed to provide a global summary of treatment-related or clinical information without imposing any correlation-based selection among predictors. All available predictors are retained, regardless of their mutual dependence structure. This formulation is particularly useful in settings where predictors capture distinct biological or clinical dimensions and are not expected to be highly correlated.

Formally, let $X = (X_1, \ldots, X_k)$ denote the full set of standardized predictors. We construct the therapeutic index using the same log-linear framework as in the composite index, but applied to the entire predictor set:

$$\log Y = a_1 \log X_1 + a_2 \log X_2 + \ldots + a_k \log X_k + e,$$

where the exponent vector $a = (a_1, \ldots a_k)$ is estimated by ordinary least squares after standardization and centering, following the same estimation procedure described in Section 2.

The resulting therapeutic index takes the multiplicative form

$$TI \propto X_1^{\hat{a}_1} X_2^{\hat{a}_2} \ldots X_k^{\hat{a}_k}.$$

Unlike the composite index, the therapeutic index does not aim to exploit redundancy among predictors. Instead, it provides an aggregate measure that reflects the joint contribution of all available variables. As a result, the therapeutic index is more broadly applicable but may be less parsimonious and potentially more sensitive to noise when predictors are weakly related.

Composite Index for Benefit-Risk Assessment [8,9]

Beyond constructing a single index, the proposed framework naturally extends to benefit–risk assessment by separating efficacy-related and safety-related information into distinct composite indices. Specifically, we first construct an efficacy index using all available efficacy parameters, such as primary and secondary clinical outcomes, biomarkers reflecting treatment benefit, or disease progression measures. All efficacy predictors are retained, and the index is estimated following the same log-linear procedure described above. Similarly, a safety index is constructed using all relevant safety parameters, which may include adverse event rates, laboratory abnormalities, drug exposure measures, and indicators of patient adherence or compliance. No correlation-based screening is imposed at this stage, allowing the safety index to capture multiple, potentially heterogeneous dimensions of treatment risk.

Formally, let $X^{(E)}$ and $X^{(S)}$ denote the sets of efficacy and safety predictors, respectively. The resulting indices take the multiplicative forms

$$fficacy\ Index \propto \prod_j (X_j^{(E)})^{\hat{a}_j^{(E)}},$$

$$Safety\ Index \propto \prod_l (X_l^{(S)})^{\hat{a}_l^{(S)}}$$

where the exponent vectors are estimated separately within each domain.

This perspective naturally supports separating efficacy-related and safety-related information into distinct yet comparable summaries, enabling transparent integration and sensitivity analyses when balancing potential gains against risks.

Concluding Remarks

Simulation

To simulate the proposed composite index procedure, we apply it to the Primary Biliary Cirrhosis (PBC) dataset. The outcome variable is the time from study registration to death, recorded as time. The candidate predictors include nine laboratory measures: bilirubin (bili), cholesterol (chol), albumin (albumin), copper (copper), alkaline phosphatase (alk.phos), aspartate aminotransferase (ast), triglycerides (trig), platelet count (platelet), and prothrombin time (protime). All analyses are conducted on complete cases. Using the penalized condition-index rule, we select the subset of predictors that balances collinearity and model complexity. A grid search over the penalty parameter λ identifies λ = 0.15 as maximizing the external reproducibility probability. The resulting subset is

$S^* = \{$ bili, ast, trig$\}$.

Applying the composite index construction to this subset yields continuous exponent estimates

$$\hat{a}_{bili} = -0.35, \hat{a}_{ast} = 0.13, \hat{a}_{trig} = -0.01.$$

Discussion

For clinical implementation, continuous exponents can be difficult to communicate or operationalize. We therefore considered a restricted set of interpretable exponents, allowing only integers and half-integers ending in .5. For each coefficient $\hat{a}_j$, we generated two candidates:

1. the nearest integer, and
2. the nearest half-integer ending in .5.

With three predictors in S^* this produces 2^3 = 8 possible rounded composite indices. For each candidate index, we re-computed the external reproducibility probability p_r and selected the rounded index that maximized p_r. The best-performing rounded index is

$$Composite\ Index_{rounded} \propto bili^{-0.5}.ast^{0.5}\ .trig^{-0.5} = \sqrt{\frac{ast}{bili\ .trig}}$$

Acknowledgments

None.

Conflict of Interest

None.

References

1. Bazett HC (1920) An analysis of the time-relations of electrocardiograms. Heart 7: 353.
2. Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons.
3. Chow SC, Lee PJ, Gao J, Lee RJ, Lee JJ, et al. (2020) Statistical method for development of composite index in clinical research. American Journal of Biomedical Science & Research 10: (4).
4. FDA (2021) Guidance for Industry - Benefit-Risk Assessment for New Drug and Biological Products. The United States Food Drug Administration, Silver Spring, Maryland.
5. Filozof C, Chow SC, Dimick-Santos L, Chen YF, Williams RN, et al. (2017) Clinical endpoints and adaptive clinical trials in precirrhotic nonalcoholic steatohepatitis: Facilitating development approaches for an emerging epidemic. Hepatology Communications 1(7): 577-585.
6. Greco S, Ishizaka A, Tasiou M, Torrisi G (2019) On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. Social Indicators Research 141: 61-94.
7. Ouellet D (2010) Benefit–risk assessment: the use of clinical utility index. Statistics in Medicine.
8. Santeramo FG (2017) Methodological challenges in building composite indexes: Linking theory to practice. In Emerging Trends in the Development and Application of Composite Indicators.
9. Zhang Y, Zhang X, Wang P, Wu YF, Chow SC (2025) A proposed confidence ellipse approach for benefit-risk assessment in clinical trials. Therapeutic Innovation & Regulatory Science 59(3): 606-618.