American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Research Article

Statistical Method for Development of Composite Index in Clinical Research

Shein-Chung Chow[1], Patty J Lee[2], Junheng Gao[1]*, Rebecca J Lee[3], Justin J Lee[4] and Ziv Soferman[5]

[1]*Department of Biostatistics and Bioinformatics, Duke University School of Medicine, North Carolina*

[2]*Allergy & Critical Care Medicine, Duke University School of Medicine, North Carolina*

[3]*Brown University, Providence, USA.*

[4]*Rutgers New Jersey Medical School, New Jersey*

[5]*Tel-Aviv Academic College of Engineering, Israel*

***Corresponding author:** Shein-Chung Chow, Patty J Lee, Junheng Gao, Duke University School of Medicine, 2424 Erwin Road, Durham, North Carolina.

To Cite This Article: *Junheng Gao, Statistical Method for Development of Composite Index in Clinical Research. 2020 - 10(4). AJBSR.MS.ID.001538.* DOI: 10.34297/AJBSR.2020.10.001538.

Received: October 06, 2020; **Published:** October 15, 2020

Abstract

In clinical research, a medical predictive modelling is often performed using a multivariate set of risk factors to predict the performance of clinical outcome for an effective disease management. Using a well-established and validated medical predictive model, our goal is to develop a composite index of several dependent predictors to better inform the disease status and/or treatment effect with more accurate and reliable assessments. In practice, since each of the multiple predictors may be positively or negatively and/or linearly or nonlinearly correlated to the clinical outcome or response, an ideal composite index should be able to account for positively/negatively and/or linearly/non-linearly associations with the clinical outcome or response. In this article, criteria and a statistical approach for development of an ideal composite index are proposed. Under the proposed criteria and procedure, statistical methods are also derived. The proposed procedure for development of the composite index is evaluated both theoretically and via a clinical trial simulation.

Keywords: Multiple Regression Analysis; Risk Factors; Medical Predictive Model; Composite Index.

Introduction

In clinical research, a medical predictive model is often established using a multivariate set of risk factors (predictors). The purpose of a medical predictive model is not only to predict the performance of clinical outcome but also to provide valuable information regarding disease management including prevention, accurate and reliable diagnosis, and effective treatment of the diseases under study. In practice, for building a medical predictive model with a multivariate set of risk factors, a (logistic) regression analysis approach is often performed by the following steps: (i) identifying potential risk factors (e.g., demographics or patient characteristics) by determining associations between the potential risk factors and the response, (ii) testing for co-linearity among the identified risk factors, (iii) performing predictive model fitting with the identified predictors, (iv) performing goodness-of-fit of the fitted model, and (v) validating the developed medical predictive model [1]. In addition, generalizability of the medical predictive model should be examined for the purpose of external validation.

A well-established and validated medical predictive model upholds sparse predictors, particularly when these predictors are highly correlated with (dependent on) one another. Therefore, the principle investigator will try to integrate multiple predictors into a single predictor that informs the disease status and/or treatment effect while upholding accuracy and reliability [2]. Since each of the multiple predictors may be positively or negatively and/or linearly

or nonlinearly correlated to the clinical outcome or response, an ideal composite index should be able to account for positive/negative and/or linear/non-linear associations with the clinical outcome.

Similar to statistical methods used for characterizing calibration (standard) curves in lab-based assays, we will develop an ideal composite index using similar criteria and corresponding methods. For example, we propose that an ideal composite index should be of the format: $x_1^a x_2^b$, where x_1 and x_2 are identified, highly correlated risk factors/predictors. Next, a procedure for the development of the ideal composite index, based on a multiple regression analysis model, is proposed. Under the multiple regression analysis model, statistical methods are derived accordingly. The criteria, process, and statistical methods are evaluated both theoretically and via a simulation study.

The remainder of this article is organized as follows: Section 2 will briefly introduce the concept of an ideal composite index, using examples from existing clinical research, and propose an innovative procedure for the development of an ideal composite index in clinical research; Section 3 will derive the statistical methods; Section 4 will apply the proposed to composite index in a simulation study.

Development of a Composite Index

Under a well-established medical predictive model, it is common to see that these predictors may be dependent on, or somehow correlated to one another in a linear/nonlinear and/or positive/negative fashion.

Ideal Composite Index

$$I_{pq}=I\_((x\_p,x\_q))= g(x_p, x_q)$$

Let $I_{pq} = I_{(x_p,x_q)} = g(x_p, x_q)$ be the composite index of x_p and x_q, where x_p and x_q are identified and highly correlated predictors which are relevant to clinical outcome and g is a utility function that combines x_p and x_q. The goal of the ideal composite index is to identify the utility function g such that the developed index can account for positive/negative and/or linear/non-linear associations between each of the predictors (i.e., x_p and x_q) and the clinical outcome.

In practice, the selection of function g depends upon the relationship between the clinical outcome and each of the predictors (i.e., x_p and x_q). For this purpose, we may consider the selection of a standard curve or calibration curve in the development and validation of an analytical method in laboratory testing. Let y be the amount of drug recovered (% of label claim) and x be the standard concentration. A standard curve or calibration curve is often determined based on the model fitting between y_i and x_i, where i=1,..,n. In assay development and validation, the following four models are commonly considered:

$$\text{Model 1: } y = \beta_0 + \beta_1 x + \varepsilon;$$
$$\text{Model 2: } y = \beta_1 x + \varepsilon;$$
$$\text{Model 3: } y = \beta_0 x^{\beta_1} \varepsilon;$$
$$\text{Model 4: } y = \beta_0 e^{\beta_1 x} \varepsilon.$$

Model 1 is linear with a non-zero intercept, while Model 2 is linear without an intercept. Model 3 and Model 4 are non-linear but can be linearized by taking the logarithm. Based on the consideration that an ideal composite index should be able to account for positive/negative and/or linear/non-linear associations between each of the predictors (i.e., x_p and x_q) and the clinical outcome (y), we propose selecting the utility function *g* as follows

$$I_{pq} = I_{(x_p,x_q)} = g(x_p, x_q) = x_p^a x_q^b$$

Example 1 – In clinical research, the above proposed composite index is commonly seen in practice. A typical example is the development of body mass index (BMI). As indicated in BMI [3], the BMI was first discussed by Quetelet in his published research work on the weight of men at different ages in 1832. BMI serves as a medical predictive model for obesity, and uses the two predictors of x_p = weight (kg) and x_q = height (m)]. In this case, BMI is given by

$$BMI = \frac{Weight}{(Height)^2} = Weight(Height)^{-2} = x_p^a x_q^b = x_p x_q^{-2}$$

where a=1 and b=-2.

Most recently, Trefethen [4] proposed a new formula for computing BMI as follows,

$$BMI = \frac{1.3Weight}{(Height)^{2.5}} = 1.3(Weight)(Height)^{-2.5} = 1.3x_p^a x_q^b = x_p x_q^{-2.5}$$

where a=1 and b=-2.5.

The scaling factor of 1.3 was determined to make the new BMI formula align with the traditional BMI formula for adults of average height. The exponent of 2.5 is a compromise between the exponent of 2 in the traditional formula for BMI and the exponent of 3 that would be expected for the scaling of weight with height.

Example 2 – Consider QT interval prolongation for cardiotoxicity. Let $x_p = QT$ interval and x_q = heart rate (RR). Bazett [5] and

Fridericia [6] proposed the following indices, which is a corrected QT interval (denoted by QT_C) adjusted for the square root of RR, respectively:

$QT_cB = \frac{QT}{\sqrt{RR}} = (QT)(RR)^{-\frac{1}{2}} = x_p^a x_q^b = x_p x_q^{-\frac{1}{2}}$, where a=1 and b=-1/2,

and

$QT_cF = \frac{QT}{\sqrt[3]{RR}} = (QT)(RR)^{-\frac{1}{3}} = x_p^a x_q^b = x_p x_q^{-\frac{1}{3}}$, where a=1 and b=-1/3.

Note that QT_CB (Bazett's index) is widely used but may give erroneous results at both slow and fast heart rates. QT_cF (Fridericia index) is also widely used and, compared to Bazett's index, may give more consistent results at fast heart rates.

Proposal for Development of a Composite Index

In the interest of minimizing predictors, especially when these predictors are dependent on one another, and upholding generality, we consider the development of a composite index based on two confirmed predictors x_p and x_q, which are corrected for each other. We propose the following steps for the development of an ideal composite index by reducing a two-parameter (x_p and x_q) problem to a single parameter (the composite index) problem.

Step 1. Establish and validate a medical predictive model. Let y denote the clinical outcome/response (independent variable) and $x_i, i = 1,....,K$ be the risk factors/predictors (dependent variables). Consider the following multiple regression model:

$$y_j = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_j, j = 1,....., n, \quad (1)$$

where $\beta_i, i = 1,....,K$ are regression coefficients and ε_i is the random error term. Under the multiple regression model, a (logistic) regression analysis approach is often performed to (i) identify potential risk factors/predictors (e.g., demographics or patient characteristics) by determining associations between the potential risk factors/predictors and the response, (ii) test for collinearity among the identified risk factors/predictors, (iii) build a medical predictive model by fitting the clinical outcome/response with the identified predictors, (iv) perform goodness-of-fit of the fitted model, and (v) validate the developed medical predictive model based on some pre-specified performance criteria.

Step 2. Under the established and validated medical predictive model, obtain estimates of the regression coefficients of the two predictors x_p and x_q, which we will develop into a composite index accordingly.

Step 3. Obtain predicted values of y based on x_p and x_q as $\hat{y}_{pqi} = \hat{\beta}_p x_{pi} + \hat{\beta}_q x_{qi}, i = 1, \ldots, n.$ Then, fit the model $\hat{y}_{pq} = x_p^a x_q^b \varepsilon$, which can be done by taking a [the?] logarithm transformation. Consequently, estimates of a and b can be obtained.

The composite index based on x_p and x_q is obtained as $CI = x_p^{\hat{a}} x_q^{\hat{b}}$.

Statistical Method

Obtain Estimates of Regression Coefficients of Predictors

Under Model (1). Without loss of generality, assume y and all x_i are standardized variables. Under standardized variables, the mean and variance of the regressors are given by

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} = 0 \text{ and } s_i^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 = 1$$

for $i = 1,....,K$. Similarly, for the standardized variable of clinical outcome response (dependent variable), we have

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = 0 \text{ and } s_y^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2 = 1.$$

Suppose we are interested in developing a composite index for two identified risk factors, namely x_p and x_q, where 1≤p≠q≤K, and x_p and x_q are known to be highly correlated in an unknown format. However, under the standardized variables, the sample covariance between x_p and x_q can be obtained as follows

$$s_{pq} = \frac{1}{n}\sum_{j=1}^{n}(x_{pj} - \bar{x}_p)(x_{qj} - \bar{x}_q) = \frac{1}{n}\sum_{j=1}^{n} x_{pj} x_{qj}$$

Similarly, sample covariance between y_i and x_{lj} is given by

$$s_{ly} = \frac{1}{n}\sum_{j=1}^{n} x_{lj} y_j$$

As a result, sample correlation between x_p and x_q and x_l and y are given by respectively.

$$r_{pq} = \frac{s_{pq}}{\sqrt{s_p^2 s_q^2}} = s_{pq} \text{ and } r_{ly} = \frac{s_{pq}}{\sqrt{s_l^2 s_y^2}} = s_{ly},$$

In order to obtain estimates of β_p and β_q, consider rewriting model (1) in the following matrix form

$$Y = X\beta + \varepsilon \quad (2)$$

where **Y** is the nx1 vector of dependent variables, **X** is the nxK matrix of regressors, **β** is the Kx1 vector of regression coefficients, and **ε** is the nx1 vector of random error terms. Under model (2), the ordinary least squares (OLS) estimator of β is given by

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

Based on standardized variables, $\hat{\beta}$ can be written as a function of their sample correlations. Denote by x_l the lth row of **X**. Thus, the (p,q)th element of $X'X$ is given by

$$(X'X)_{pq} = (\sum_{j=1}^{n} x_{j.}' x_{j.})_{pq} = \sum_{j=1}^{n} x_{pj} x_{qj} = ns_{pq} = n_{pq}$$

Furthermore, the p[th] element of $X'Y$ is

$$(X'Y)_p = (\sum_{j=1}^{n} x_{j.}' Y)_p = \sum_{j=1}^{n} x_{pj} y_j = ns_{py} = nr_{py}$$

Now, denote by r_{xx} the sample correlation matrix of X. That is, the KxK matrix whose (p,q) entry is equal to r_{pq}. Thus, $X'X = nr_{xx}$. Similarly, denote by r_{XY} the $Kx1$ vector whose p^{th} entry is equal to r_{PY}. Thus, $X'Y = nr_{XY}$. This implies

$$\hat{\beta} = (X'X)^{-1}X'Y = r_{XX}^{-1} r_{XY} \quad (3)$$

The estimates of β_p and β_q are the p^{th} and q^{th} elements of $\hat{\beta}$

Fitting the Model between Predicted Values and the Composite Index

For the development of the composite index based on x_p and x_q, which are dependent on each other,

consider fitting the following multicaptive model between the predicted values

$\hat{y}_{pqi} = \hat{\beta}_p x_{pi} + \hat{\beta}_q x_{qi}, i = 1,.....,n$ and the composite variable, $x_p^a x_q^b$:

$$\hat{y}_{pq} = x_p^a x_q^b \varepsilon,$$

or $\log(\hat{y}_{pq}) = a\log(x_p) + b\log(x_q) + \log(\varepsilon),$ which is equivalent to

$$\hat{y}_{pq}' = ax_p' + bx_q' + e,$$

where

$$\hat{y}_{pq}' = \log(\hat{y}_{pq}), x_p' = \log(x_p), x_q' = \log(x_q) \text{ and } e = \log(\varepsilon).$$

Following the idea described in Section 3.1, estimates of a and b can be similarly obtained.

Validation of the Developed Composite Index

We may validate the developed composite index by considering how close an observed y, its predicted value $\hat{y}_{pq}$ (obtained from $\hat{y}_{pq} = \hat{\beta}_p x_p + \hat{\beta}_q x_q$) and the predicted values $\tilde{y}_{pq}$ (obtained from $\tilde{y}_{pq} = x_p^{\hat{a}} x_q^{\hat{b}}$ based on the fitted regression model discussed in the previous sections) are to one another. To assess the closeness, we propose the following two measures, which are based on either the absolute difference or the relative difference between $\hat{y}_{pq}$ and $\tilde{y}_{pq}$:

Criterion I. $p_1 = P\{|\hat{y}_{pq} - \tilde{y}_{pq}| < \delta\},$

Criterion II. $p_2 = P\left\{\left|\frac{\hat{y}_{pq} - \tilde{y}_{pq}}{\hat{y}_{pq}}\right| < \delta\right\}.$

In other words, it is desirable to have a high probability that the difference or the relative difference between $\hat{y}_{pq}$ and $\tilde{y}_{pq}$, given by p_1 and p_2 respectively, is less than a clinically or scientifically meaningful δ. Then, for either i = 1 or 2, it is of interest to test the following hypotheses

$$H_0 : p_i \le p_0 \text{ versus } H_a : p_i > p_0, \quad (4)$$

where p_0 is some pre-specified constant. If the conclusion is to reject H_0 in favor H_a, then the developed composite index is considered validated. The technical details of the test of hypothesis corresponding to the two criteria can be found in Tse et al. [7].

Simulation Study

In the simulation study, we construct a composite index that follows the similar modelling of body mass index with randomly generated weight x_p and height x_q data. We examine three scenarios at which predictive accuracy could be affected. Firstly, we run the simulation for different sample sizes and compare the closeness of developed composite index to the observed value. Secondly, we analyse whether the correlation relationship between the predictors will affect the predictive accuracy. Thirdly, we test under different regression equations for observed y_i and x_i, the estimation of composite index $\tilde{y}_{pq}$ is still consistent to the predicted values $\hat{y}_{pq}$.

In each case, we generated observed values for weight x_p and height x_q from bivariate normal distribution with correlation ρ =0.6. The mean of weight and height are 70 kg and 1.8m and variance are 15 and 0.15 respectively. Those values refer to the BMI historical data reported in the National Health Examination surveys. We assumed the true relationship between the clinical outcome y_{pq} and the risk factors to be $y = 52 + 0.4 \cdot x_{weight} + 33 \cdot x_{height}$. We fitted the relationship between y and x_{weight}, x_{height} by linear regression, and obtained the coefficient estimates and predictive outcome $\hat{y} = \hat{\beta}_p \cdot x_{weight} + \hat{\beta}_q \cdot x_{height}$. Then we applied the linear model for $\log(\tilde{y}) = a\log(x_{weight}) + b\log(x_{height})$ and obtained the composite index as $\tilde{y} = x_{weight}^{\hat{a}} \cdot x_{height}^{\hat{b}}$. Here our composite index is analogous to the widely used BMI formulation. To test the predictive power of the composite index, we run the simulation for 1000 times. The differences between $\hat{y}_{pq}$ and $\tilde{y}_{pq}$ under different sample sizes, correlation coefficients, and combinations of β coefficients were summarized in Table 1.

n	$\lvert\hat{y}_{pq} - \tilde{y}_{pq}\rvert$	$\left\lvert\frac{\hat{y}_{pq} - \tilde{y}_{pq}}{\hat{y}_{pq}}\right\rvert$	$\hat{a}$	$\hat{b}$
50	1.012	0.040	1.034	-2.050
100	1.064	0.042	1.032	-2.041
500	1.092	0.043	1.032	-2.035
1000	1.098	0.043	1.034	-2.036
10000	1.100	0.043	1.033	-2.036

Scenario I: For different sample size n = 50, 100, 500, 1000, 10000.

ρ	$\lvert\hat{y}_{pq} - \tilde{y}_{pq}\rvert$	$\left\lvert\frac{\hat{y}_{pq} - \tilde{y}_{pq}}{\hat{y}_{pq}}\right\rvert$	$\hat{a}$	$\hat{b}$
-0.9	1.005	0.039	1.219	-2.041
-0.7	1.021	0.040	1.132	-2.056
-0.5	1.034	0.040	1.104	-2.058
-0.3	1.046	0.041	1.089	-2.056
-0.1	1.057	0.041	1.077	-2.052
0.1	1.069	0.042	1.068	-2.049
0.3	1.081	0.042	1.055	-2.044
0.5	1.090	0.043	1.044	-2.04
0.7	1.105	0.043	1.017	-2.031
0.9	1.111	0.044	0.915	-2.006

Scenario II: For different correlation ρ from -0.9 to 0.9 by 0.2 (n = 1000).

β_p, β_q	$\lvert\hat{y}_{pq} - \tilde{y}_{pq}\rvert$	$\left\lvert\frac{\hat{y}_{pq} - \tilde{y}_{pq}}{\hat{y}_{pq}}\right\rvert$	$\hat{a}$	$\hat{b}$
(1, 1)	0.049	0.001	0.763	0.018
(1, 5)	0.178	0.002	0.714	0.083
(1, 10)	0.324	0.003	0.661	0.153
(1, 20)	0.564	0.004	0.577	0.265
(2, -1)	0.039	0.000	0.883	-0.010
(2, -5)	0.188	0.001	0.921	-0.054
(2, -10)	0.398	0.003	0.973	-0.116
(2, -20)	0.902	0.007	1.099	-0.264

Scenario III: For different beta coefficients (n=1000).

Table 1: Evaluation of the Closeness between $\hat{y}_{pq} - \tilde{y}_{pq}$.

As shown in Table 1, the estimation of the proposed composite index is not affected by the sample size. Even for small sample size such as n = 50, the estimation is accurate and consistent. In scenario II, the difference between $\hat{y}_{pq}$ and $\tilde{y}_{pq}$ is minor for proposed correlation between x_p and x_q. With higher correlation coefficient, the difference becomes relatively smaller. For different beta coefficients, the relative difference is robust while the absolute difference has a minor trend of increase when response variable

increases. Our results suggest the need to scale the risk factors or leave aside the intercept term before applying the linear regression model for the log-transformed variables.

Concluding Remarks

In this article, although the development of the composite index focuses on two dependent predictors, similar idea can be easily extended to develop a composite index combining more than two predictors, which are correlated one another. In the multiple regression analysis, the ordinary least squares (OLS) approach is considered for obtaining the estimates of regression coefficients. In practice, alternatively, we may consider weighted OLS (WOLS) for adjustment if heterogeneity in predictors are present.

For simplicity and illustration purpose, we consider the predictors $x_i, i = 1, ..., K$ are of the same data type such as continuous variable. In practice, however, $x_i, i = 1, ..., K$ may be of different data types (e.g., continuous, binary response, or time-to-event data). In this case, the idea regarding the use of standardized variables as discussed in Chow and Huang [8] may be similarly applied.

An ideal composite index should possess 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.

It should be noted that the developed composite index depends upon estimates of a and b, i.e., $\hat{a}$ and $\hat{b}$, which may not be integers. In this case, we may consider using $[\hat{a}]$ and $[\hat{b}]$, where [k] denotes the nearest integer of k. This may explain why Trefethen [8] obtained a "-2.5" rather than "-2" for development of BMI. In practice, however, it is suggested that the adjacent nearest integers be evaluated for selection of optimal composite index [9].

A future possibility is to build and check the quality of a composite index containing a variable for body structure and a variable for muscle mass, especially since risk for diabetes can be influenced by relative muscle mass for people with same height and weight, thus same BMI. In addition to body frame, such as large or small, despite same height could also be taken into account.

References

1. Liu M, Chow SC (2018) Logistic regression process: predictive model building in clinical research. In Encyclopedia of Biopharmaceutical Statistics, 4th Edition, Ed. Chow, SC, CRC Press, Taylor & Francis, New York. pp. 1309-1314.
2. Chow SC, Lee PJ (2020) Time to revisit endpoint selection in clinical trials. American Journal of Biomedical Science & Research 9(3).
3. BMI (Body Mass Index) (2013).
4. Trefethen N (2019) New BMI (Body Mass Index). Mathematical Institute, University of Oxford, USA.
5. Bazett HC (1920) An analysis of the time-relations of electrocardiograms. Heart 7:353.
6. Fridericia LS (1920) Die systolendaur im elektrokardiogram bei normalen menschen und bei herzkrauken. Acta Med Scand 53:469.
7. Tse SK, Chow SC, Yang C (2008) Statistical tests for one-way/two-way translation in translational medicine. Journal of Formosan Medical Association, 107: 12, S42-S50.
8. Chow SC, Huang Z (2019) Innovative thinking on endpoint selection in clinical trials. Journal of Biopharmaceutical Statistics 29(5): 941-951.
9. Chow SC, Liu JP (1995) Statistical Design and Analysis in Pharmaceutical Science – Validation, Process Controls, and Stability. Marcel Dekker, New York.