DOI: 10.7860/JCDR/2022/55379.16629

Original Article

Statistics Section

# Application of Principal Component Analysis in Dealing with Multicollinearity in Modelling Clinical Data

AKASH MISHRA[1], N SREEKUMARAN NAIR[2], KT HARICHANDRAKUMAR[3], VS BINU[4], SANTHOSH SATHEESH[5]




## ABSTRACT

**Introduction:** One of the stringent assumptions about covariates in the Cox hazard and Logistic regression modelling is that they should be independent. Incorporating correlated covariates as such into the model might distort the precision of the estimates due to multicollinearity. One way to deal with multicollinearity is by using Principal Component Analysis (PCA) technique.

**Aim:** To demonstrate the application of PCA in dealing with correlated covariates while modelling time to event and case-control study data.

**Materials and Methods:** This study was conducted at Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, India, from February 2021 to January 2022. Two datasets were used for the demonstration i.e., data relates to a time to event outcome and a case-control study with binary outcome in which lipids were the correlated covariates. Three sets of Cox regression models were used to demonstrate change in hazard ratios with 95% Confidence Intervals (CI) for evaluating the effect of intervention at a different time of lipid measurement. Model I has evaluated treatment/Body Mass Index (BMI) effect on the outcome by ignoring the effect of lipid parameters. Model II has evaluated treatment/BMI effect on the outcome by incorporating lipid variables but ignoring multicollinearity. Model III has evaluated treatment/BMI effect on the outcome by incorporating lipid variables through principal component analysis and thus adjusting for multicollinearity. Similarly, a logistic regression model was performed by using the same three sets of models to evaluate the effect of exposure (BMI). The comparability of lipids between the two groups for both datasets was tested using Hotelling's T-squared statistic.

**Results:** The lipids measured at 12$^{th}$, 24$^{th}$ and 36$^{th}$ months between the two groups in the first data set as well as between cases and controls in the second data set were statistically significant. In the first dataset, at baseline, the Hazard Ratio's (HR's) were statistically similar irrespective of the models used; while decreasing successively with narrowing 95% CI's as moving from model I to model III for the lipid measured at 12$^{th}$, 24$^{th}$ and 36$^{th}$ months. Further, at 24$^{th}$ and 36$^{th}$ months, the HR in model-III found to be significant. In the second data set, the Odds Ratio (OR) were significant for all the three models and it was almost similar for model I and II but in model III it was elevated.

**Conclusion:** The multicollinearity issue should be properly addressed before including correlated covariates in the Cox regression hazard and Logistic regression model. The PCA technique would be a favourable method.


**Keywords:** Analysis of correlated outcomes, Case-control study, Cox 36 regression hazard model, Logistic regression model

## INTRODUCTION

Any model that establishes the effect of the potential covariates on the outcome variable should comply with the nature of the outcome or dependent variable. For a longitudinal study with a time to event outcome variable, the commonly used statistical approach is the Cox hazard regression model [1,2]. Similarly, for a case-control study with a binary outcome variable, the suitable approach is the logistic regression model [1]. One of the assumptions of such a regression model is that the predictor variables should not be correlated with each other. However, the predictors under consideration may not be truly independent but rather correlated in biomedical research. Such dependency between the covariates in the regression modelling leads to a condition referred to in a statistical term as multicollinearity which means a covariate can be predicted by the remaining covariates [3,4].

The main issue with multicollinearity is that the estimate of the regression coefficient of one of the correlated predictors depends on the presence of the other predictors in the model. Also, due to multicollinearity, the estimated standard errors of the regression coefficients might get inflated and could lead to spurious results. Variables in clinical research studies are usually found to be correlated [5-9]. This stipulates that the change in one variable is associated with the change in another variable. There are studies that have established the association of these lipid parameters with the outcome of interest such as Cardiovascular Disease (CVD) and Sudden Sensorineural Hearing Loss (SSNHL) [10-24].

The researcher while evaluating the effect of an intervention/exposure on the outcome has to be conscious in dealing with the effect of such multiple correlated predictors. There are studies where the multicollinearity issues of lipids were not addressed while modelling the outcome variable with Cox hazard or logistic regression models [19-24]. In these cited studies, the effect of intervention/exposure were evaluated by introducing the lipid parameters as the covariates in the model as such. So, due to multicollinearity, it is likely to get unreliable point estimates of Hazard Ratio (HR) or Odds Ratio (OR) of the intervention/exposure. Moreover, incorporating the correlated covariates into the model as such may weakens the statistical power of such regression models. In such conditions, the researcher will be concluding with a compromised precision of the effect of intervention/exposure. To address the multicollinearity issue, methods like partial least square (PLS), Ridge Regression (RR) and Principal Component Analysis (PCA) have been suggested. PLS and RR methods are used for the continuous outcome variable. Since our outcome variable is binary, so, in this article PCA technique was used [25].

The objective of this study was to demonstrate the application of PCA method in dealing with multicollinearity with Cox and logistic regression models. The demonstration was done from two data sets. The first data set was from the ACCORD BP (Action to Control Cardiovascular Risk in Diabetes Blood Pressure) trial in which data was recorded from time to event. While second data set was

from a case-control study on Sudden Sensorineural Hearing Loss (SSNHL). Lipids were then correlated with covariates in both the data sets.

## MATERIALS AND METHODS

This study was conducted at Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, India, from February 2021 to January 2022.

### Brief Description of Dataset

For the demonstration the following two datasets were used.

**ACCORD BP trial dataset [26]:** The ACCORD trial dataset was available from Biologic Specimen and Data Repository Information Coordinating Centre (https://biolincc.nhlbi.nih.gov/home/) of National Heart, Lung, and Blood Institute, upon institutional request. It was an open-label multicentric randomised trial of 84 months follow-up. A total of 4733 high-risk type 2 diabetes mellitus eligible participants were randomised into two study groups:

- Intensive BP control group (n=2362): Treatment strategy was to lower Systolic Blood Pressure (SBP) below 120 mmHg.
- Standard BP control group (n=2371): The strategy was to lower SBP below 140 mmHg.

The treatment strategy followed in the respective BP control groups was for the comparison in reducing CVD events. The primary outcome variable considered was a composite of non fatal Myocardial Infarction (MI), non fatal stroke and CVD death whichever occurred first.

The five lipid parameters were measured at baseline and thereafter on yearly basis:

- Total Cholesterol (TC)
- Triglyceride (TG)
- Very Low Density Lipoprotein (VLDL)
- Low Density Lipoprotein (LDL)
- High Density Lipoprotein (HDL)

The participants who were not measured for their lipid parameters at different follow-ups were excluded from the analysis.

**SSNHL case-control study dataset [10]:** The SSNHL case-control study dataset was publicly available from the authors obtained by dryad (http://dx.doi.org/10.5061/dryad.r2b1n). A total of 324 hospitalised cases for SSNHL and 972 controls with normal hearing were taken. As per World Health Organisation (WHO) criteria the underweight subject (BMI ≤18.5 kg/m$^2$) from among the cases and controls were excluded from the analysis [27]. The data on BMI and lipid parameters TC, TG, LDL and HDL for the cases and controls were available.

### Models

**Cox hazard regression model:** The Cox proportional hazard model was used when the covariates considered in the model satisfied the proportionality assumption. For a random binary outcome variable Y with a vector of covariates X: [$X_1$ $X_2$ .....$X_p$] and the corresponding vector of β coefficients β'=[$\beta_1$ $\beta_2$.....$\beta_p$], Cox proportional hazard model with hazard rate h (t/X) at any time t is expressed as:

$h(t/X)=h_0(t)\, e^{X\beta}$

Where, $h_0$ (t) is an unspecified non negative function of time called baseline hazard at time t. Thus, the HR to an individual of $j^{th}$ group with 1×p vector of covariates X against an individual of $k^{th}$ group with a vector of same covariates can be obtained as:

$$HR = \frac{h(t/\mathbf{X}_j)}{h(t/\mathbf{X}_k)} = e^{(\mathbf{X}_j - \mathbf{X}_K)\boldsymbol{\beta}}$$

Where, $X_j$ and $X_k$ are the vector of the same covariate X for $j^{th}$ and $k^{th}$ groups respectively [1,3]:

**Cox time-dependent hazard model:** The Cox time-dependent model was used when the covariates considered did not satisfy the proportionality assumption. The time-dependent Cox hazard model with hazard rate h[t/X(t)] at time t is expressed as:

$h[(t/X(t)]=h_0(t)\, e^{X(t)\beta}$

Where, $h_0$ (t) is an unspecified non negative function of time called base line hazard at time t. Thus, the HR can be obtained as:

$$HR = \frac{h[(t/\mathbf{X}_j(t)]}{h[(t/\mathbf{X}_k(t)} = e^{[(\mathbf{X}_j(t) - \mathbf{X}_K(t)]\boldsymbol{\beta}}$$

Where, $X_j$ (t) and $X_k$ (t) are the vector of the same covariate X(t) at time t for the $j^{th}$ and $k^{th}$ groups respectively. The estimate of HR associated with $i^{th}$ covariate and the corresponding (1-α) Confidence Interval (CI) was obtained by using the estimates of $\beta_i$ and its standard error as $e^{\beta_i}$ and $e^{\beta_i + Z_{\alpha/2}\, SE(\beta_i)}$, respectively.

**Logistic regression model:** The logistic regression model was used in case control study data set. For a random outcome variable Y with vector of covariates X: $X_1$ $X_2$ .....$X_p$ and the corresponding vector of β coefficients β'=[$\beta_0$ $\beta_1$ $\beta_2$.....$\beta_p$], the estimate of odds ratio (OR) was obtained by using the logistic model as:

$$p = P[Y = 1/\boldsymbol{X}] = \frac{1}{[1 + e^{-X\beta}]}$$

Where, $\beta_0$ is the constant, called intercept of the regression equation.

Thus, the odds (OR) to an individual of $j^{th}$ group with $p_j$ being the probability of occurrence of event with vector of the covariates X against an individual of $k^{th}$ group with $p_k$ being the probability of occurrence of the event with the vector of same covariates can be obtained as:

$$OR = \frac{[p_j/(1-p_j)]}{[p_k/(1-p_k)]} = e^{(X_j - X_K)\boldsymbol{\beta}}$$

The estimates of OR associated with $i^{th}$ covariate and the corresponding (1-α) CI was obtained by using the estimates of $\beta_i$ and its standard error as $e^{\beta_i}$ and $e^{\beta_i + Z_{\alpha/2}}$, respectively.

**Principal Component Analysis (PCA):** It is a data dimension reduction technique. It creates a new set of uncorrelated variables known as Principal Components (PC) based on the linear combinations of all correlated variables. Generally, first few PC's can explain the most of total variability of all correlated variables [28].

The general PCA equation to create the independent variables is given by:

$$PC_i = \mathbf{e}_i^T\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p \quad i = 1, 2, \ldots p$$

which maximises the variance of ($e^i_T$ X) subject to the condition $e^i_T$ $e_i$=1 and Cov ($e_i^T$ X, $e_k^T$X)=0 for i≠k, where, X'=[$X_1$ $X_2$ $X_3$…$X_p$], a random vector of correlated p variables which have the covariance matrix as Σ with the eigen values $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0$ and $e_i^T$ is the transpose of eigen vector corresponding to $i^{th}$ eigen value ($\lambda_i$). All the PC's are uncorrelated and variance equal to the eigen values of Σ i.e., Var ($PC_i$) = $\lambda_i$. Thus, the first PC explains the maximum variation of the data followed by second component and so on. For both the data set, the new independent variables were created using the measured values of lipid parameters.

## STATISTICAL ANALYSIS

The following three sets of models were used in analysis for both the data sets:

**Model I:** Treatment/BMI effect on the outcome been compared by ignoring the effect of lipid parameters.

**Model II:** Treatment/BMI effect on the outcome been compared by incorporating lipid variables but ignoring multicollinearity.

**Model III:** Treatment/BMI effect on the outcome has been compared by incorporating lipid variables through principal component analysis and thus adjusting for multicollinearity.

However, the methodological component and analysis part were explained for each dataset separately.

**ACCORD BP trial dataset:** The Pearson correlation coefficients between lipids parameters were computed with log-transformed values of lipids due to their skewed distribution [Table/Fig-1].

Three Cox proportional hazard regression models were fitted. The treatment group was taken as the main predictor variable. The Cox proportional hazard regression model was used if proportionality assumptions were satisfied; the Cox time-dependent regression hazard model otherwise. The proportionality assumption of each covariate was tested by Schoenfeld's global test [29].

The HR with 95% CIs were estimated across all the above three models for measurements of lipids at the baseline, 12th, 24th and 36th month follow-ups. The lipid parameters were introduced after seeing the significant difference in lipids between the two treatment groups. This was tested by multivariate Hotelling's T-squared statistic as the lipids were correlated [28,30]. The difference testing performed on log-transformed values of lipids for both the datasets to meet assumptions as the distributions were skewed. The lipid parameters were found to differ significantly between the two groups at each time point except at baseline [Table/Fig-2].

The eigen values ($\lambda_i$) and the corresponding transpose of eigen vectors ($e_i^T$) were obtained for intensive and standard BP control groups separately. Further, PCA was performed in each group to create new independent variables for the random vector X'=[TC,TG,VLDL,LDL,HDL] with the covariance matrix as:

$$\Sigma_{ACCORD} = \begin{bmatrix} \sigma_{TC} & \sigma_{TC,TG} & \sigma_{TC,VLDL} & \sigma_{TC,LDL} & \sigma_{TC,HDL} \\ \sigma_{TG,TC} & \sigma_{TG} & \sigma_{TG,VLDL} & \sigma_{TG,LDL} & \sigma_{TG,HDL} \\ \sigma_{VLDL,TC} & \sigma_{VLDL,TG} & \sigma_{VLDL} & \sigma_{VLDL,LDL} & \sigma_{VLDL,HDL} \\ \sigma_{LDL,TC} & \sigma_{LDL,TG} & \sigma_{LDL,VLDL} & \sigma_{LDL} & \sigma_{LDL,HDL} \\ \sigma_{HDL,TC} & \sigma_{HDL,TG} & \sigma_{HDL,VLDL} & \sigma_{HDL,LDL} & \sigma_{HDL} \end{bmatrix}$$

Using PC equations, data was generated for the first three independent PC's at baseline, 12th, 24th and 36th months, respectively. These first three PC's were able to explain more than 99% of the total variation in lipids at each considered time point. The effect of intervention in model III was evaluated by adjusting for the effect of newly formed independent PC's in the Cox hazard model. The significance of HR's was judged by their 95% CI's.

**SSNHL case control study dataset:** Similarly, the correlation coefficients between lipids parameters were computed with log transformed values of lipids [Table/Fig-1]. The BMI was considered as the primary exposure for the SSNHL data. The same three sets of Logistic regression models were fitted for SSNHL dataset. The OR with 95% CI were estimated for BMI which was categorised as normal (BMI between 18.5 to 24.99 kg/m$^2$) and overweight or obese (BMI ≥25 kg/m$^2$) [29]. Again, the difference in the lipid parameters between cases and controls was tested using multivariate Hotelling's T-squared statistic [Table/Fig-2]. The lipids were introduced into the model as these differed significantly between the groups.

Similarly, the eigen values ($\lambda_i$) and the corresponding transpose of eigen vectors ($e_i^T$) were obtained for cases and control separately and corresponding PCA was performed for each group for a random vector X'=[TC,TG,LDL,HDL] with covariance matrix as:

$$\Sigma_{SSNHL} = \begin{bmatrix} \sigma_{TC} & \sigma_{TC,TG} & \sigma_{TC,LDL} & \sigma_{TC,HDL} \\ \sigma_{TG,TC} & \sigma_{TG} & \sigma_{TG,LDL} & \sigma_{TG,HDL} \\ \sigma_{LDL,TC} & \sigma_{LDL,TG} & \sigma_{LDL} & \sigma_{LDL,HDL} \\ \sigma_{HDL,TC} & \sigma_{HDL,TG} & \sigma_{HDL,LDL} & \sigma_{HDL} \end{bmatrix}$$

The first three independent PC's were generated, which were able to explain 99% of total variation of all lipids. The effect of BMI in model III was evaluated by adjusting for the effect of newly formed

| | Accord BP trial dataset | | | | | SSNHL dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Lipids | TC | TG | LDL | HDL | VLDL | Lipids | TC | TG | LDL | HDL |
| TC | 1 | 0.377** | 0.807** | 0.132** | 0.389** | TC | 1 | 0.211** | 0.908** | 0.293** |
| TG | | 1 | -0.043** | -0.574** | 0.990** | TG | | 1 | -0.032 | -0.365** |
| LDL | | | 1 | 0.186** | -0.025 | LDL | | | 1 | 0.062* |
| HDL | | | | 1 | 0.565** | HDL | | | | 1 |
| VLDL | | | | | 1 | | | | | |

**[Table/Fig-1]:** Correlation between the lipid parameters.
*,**Correlation is significant at 0.05 and 0.01 level of significance(2-tailed)
TC: Total cholesterol; TG: Triglyceride; VLDL: Very low-density lipoprotein; LDL: Low density lipoprotein and HDL: High-density lipoprotein; SSNHL: Sudden sensorineural hearing loss

| Time of lipid measurement | Bloop pressure group | Number | ACCORD BP trial dataset Lipid Parameters (Mean±SD) | | | | | Multivariate approach* |
|---|---|---|---|---|---|---|---|---|
| | | | TC | TG | LDL | HDL | VLDL | p-value |
| Baseline | Intensive | 2355 | 194.09± 45.07 | 194.93±177.77 | 111.11±37.39 | 46.08± 3.21 | 36.90±28.97 | 0.262 |
| | Standard | 2359 | 191.43±44.36 | 191.26±184.63 | 108.78±35.98 | 46.42±14.09 | 36.22±28.55 | |
| 12th month | Intensive | 1986 | 194.28±52.12 | 201.53±183.51 | 109.98±40.38 | 45.40±13.36 | 38.90±32.96 | 0.002 |
| | Standard | 1969 | 189.76±48.78 | 184.49±165.04 | 107.39±37.70 | 46.84±14.22 | 35.52±29.29 | |
| 24th month | Intensive | 2131 | 188.21±49.61 | 189.34±171.47 | 105.92±40.38 | 45.88±13.39 | 36.46±28.91 | 0.022 |
| | Standard | 2154 | 185.53±45.89 | 174.11±134.92 | 104.76±37.81 | 46.99±14.17 | 33.78±23.66 | |
| 36th month | Intensive | 2015 | 181.70±48.15 | 185.07±177.37 | 100.76±39.40 | 45.32±13.53 | 35.61±27.05 | 0.007 |
| | Standard | 2077 | 179.07±44.01 | 169.58±145.22 | 99.42±36.35 | 46.80±14.37 | 32.86±23.64 | |
| Sudden sensorineural hearing loss dataset | | | | | | | | |
| Group | | Number | | Lipid Parameters (Mean±SD) | | | | Multivariate approach* |
| | | | TC | TG | LDL | HDL | | p-value |
| Cases | | 312 | 193.62±39.00 | 123.67±78.10 | 111.21±46.86 | 57.71±15.39 | | <0.001 |
| Controls | | 927 | 184.31±35.17 | 111.42±66.07 | 108.02±31.33 | 54.03±13.01 | | |

**[Table/Fig-2]:** Summary statistics of lipid parameters and decision about difference in lipid parameters between two groups for each data sets.
*Comparisons were done on log transformed values.
BP Group: Blood pressure group; TC: Total cholesterol; TG: Triglyceride; VLDL: Very low-density lipoprotein; LDL: Low density lipoprotein and HDL: High-density lipoprotein; SD: Standard deviation; SSNHL: Sudden sensorineural hearing loss dataset

independent PC's into logistic regression model. The significance of OR's was judged by their 95% CI's.

The analysis was carried using R Studio version 3.6.1 [31], Statistical Package for Social Sciences (SPSS) version 19.0 [32] and StataCorp. volume 13 [33].

## RESULTS

The correlation coefficient and their significance between the lipid parameters were shown in [Table/Fig-1] for both datasets. For the ACCORD BP trial data, the intensive and standard BP control groups at baseline were statistically similar for lipid parameters but differed significantly at 12th, 24th and 36th months [Table/Fig-2]. Similarly, the lipid profiles of cases and control groups were significantly different in the SSNHL data [Table/Fig-2].

For the ACCORD BP trial data, [Table/Fig-3] gives the comparison of effect of intervention (HR and 95% CI) assessed by three different models at the different time of lipid measurements. At baseline, the HR's were statistically similar with slight variation irrespective of the models used. While at 12th, 24th and 36th months, the scenario of HR's was different in the three models. The HR's were successively decreasing with narrowing 95% CI's as moving from model-I to model-III. For the lipid measurement at 12th month, the HR's and the corresponding 95% CI in the successive three models were 0.885 (0.713-1.099), 0.860 (0.692-1.068), 0.835 (0.672-1.038). At 24th month the HR's and the corresponding 95% CI in the respective models were 0.835 (0.681-1.025), 0.818 (0.667-1.004), 0.806 (0.657-0.990) and at 36th month the HR's and 95% CI for the three models were 0.835 (0.674-1.035), 0.820 (0.656-1.025), 0.695 (0.546-0.883). Moreover, at 24th and 36th months, the HR with model-I and model-II were insignificant but found to be significant for the model-III.

| Time of lipid measurement | Model I: Effect of intervention by ignoring lipids | Model II: Effect of intervention by including lipid as covariates but ignoring multicollinearity | Model III Effect of intervention by including lipid through PCA |
|---|---|---|---|
| | HR (95% CI) | HR (95% CI) | HR (95% CI) |
| Baseline | 0.891 (0.739-1.074) | 0.885 (0.734-1.067) | 0.918 (0.760-1.108) |
| 12th month | 0.885 (0.713-1.099) | 0.860 (0.692-1.068) | 0.835 (0.672-1.038) |
| 24th month | 0.835 (0.681-1.025) | 0.818(0.667-1.004) | 0.806 (0.657-0.990) |
| 36th month | 0.835 (0.674-1.035) | 0.820 (0.656-1.025) | 0.695 (0.546-0.883) |

**[Table/Fig-3]:** Effect of intervention (Hazard Ratios and 95% CI) assessed by three models at different time of lipid measurements from ACCORD BP Trial dataset.
HR: Hazard ratio; 95 % CI: 95% Confidence interval

For SSNHL data the effect of BMI (OR and 95% CI) was compared between the three models [Table/Fig-4]. The OR with 95% CI for models I and II was 1.465 (1.119-1.917) and 1.467 (1.106-1.945), respectively which indicates the similarity of the point estimate and the corresponding precisions did not differ much. But, model III showed a different scenario as compared to models I and II. The OR with 95% CI for model III was 1.988 (1.425-2.773). The OR for model III was relatively elevated as compared to models I and II and 95% CI was wider too.

| Model I (Effect of BMI by ignoring role of lipids) | Model II (Effect of BMI by including lipid as covariates but ignoring multicollinearity) | Model III (Effect of BMI by including lipid through PCA) |
|---|---|---|
| OR (95% CI) | OR (95% CI) | OR (95% CI) |
| 1.465 (1.119-1.917) | 1.467 (1.106-1.945) | 1.988 (1.425-2.773) |

**[Table/Fig-4]:** Effect of BMI (Odds Ratios and 95% CI) assessed by three different models from SSNHL dataset.
BMI: Body mass index; OR: Odds ratio; 95% CI: 95% Confidence interval

## DISCUSSION

In Cox proportional hazard and logistic regression models, the multicollinearity assumptions on the covariates often get oversighted in medical research. This leads to compromised precision of the estimates. Such covariates needed to be independent when considered in the model. Otherwise, the presence of multicollinearity may distort true estimates and thus, end up with biased findings [1,3,4]. The multivariate statistical approach which deals with the multiple correlated outcomes has its own applications to deal with such problems. The PCA is one of them which has the potential to derive independent PC's. Moreover, it reduces the dimension of the correlated data and the first few components can explain almost total variation in the data. Thus, instead of using the correlated covariates as such in the model, a few PC's can be included in the model without loss of information. This PCA approach addresses the issue of multicollinearity with a smaller number of predictors.

There are few cited studies that had evaluated the effect of interventions/exposure on the outcome. In these studies, the lipids parameters that are associated with the outcome of interest were incorporated into the model as such. Thus, by ignoring the effect of multicollinearity conclusions were made [19-24]. Pedersen TR et al., used the Cox hazard model to compare the event rate of the primary outcome of major coronary events in patients treated with high-dose of atorvastatin against usual-dose. The HR was estimated for the primary endpoint adjusting for the other variables including TC and HDL as the simultaneous covariates. The decision emerged in support of the high dose of Atorvastatin in reducing the primary outcome [19]. However, the precision of the estimated HR would have been more reliable, if the multicollinearity among the lipids would have been addressed using PCA. Ting ZWR et al., examined the effects of the use of statins and fibrates on the onset of CVD in Chinese diabetic patients using the Cox model. The HR's were estimated for the lipids LDL, HDL and TG by adjusting the effects of several identified covariates. These correlated lipids had been considered as the separate covariates in the model. The reliability of the estimates may be questionable as the multicollinearity among them was ignored. [20]. Hou Q et al., by using a logistic regression model identified the relevant predictors of the presence of carotid plaque in the general Chinese adults. They identified age, gender, DBP and TC as the independent predictors of carotid plaque. Since, age, DBP and TC are the correlated predictors, the estimates of OR's of these as well as of gender may not be precise as they did not account for multicollinearity. Atleast by using PCA, the more precise estimate for gender could have been obtained [22]. The present study demonstrated the application of PCA technique in dealing with multiple correlated covariates. This could benefit the medical researchers/clinicians to obtain more valid and precise estimates for the effect of intervention/exposure. The findings of the ACCORD BP trial data set and SSNHL case-control study dataset for all the three comparative models suggest the importance of PCA to enhance the reliability of the estimates with improved precision [Table/Fig-3]. Although, this study demonstrated the application of PCA to address multicollinearity for continuous correlated covariates. But this concept could be employed for correlated categorical covariates also using PCA technique. It could be a good motivation and an interesting area of future research.

### Limitation(s)

This study demonstrated the application of PCA to address multicollinearity for continuous correlated covariates and not for categorical correlated covariates.

## CONCLUSION(S)

The study clearly demonstrates that multicollinearity among the covariates in the model should be addressed before inclusion in the Cox regression or Logistic regression model. The PCA technique could be one of the ways to address this issue to obtain reliable and precise estimates for the covariates of interest.

## Acknowledgement

Authors are sincerely grateful to ACCORD Research Materials obtained from National Heart, Lung, and Blood Institute (NHLBI), Biologic Specimen and Data Repository Information Coordinating Centre for providing access to their data through Research Materials Distribution Agreement (RMDA) and authors are also thankful to the ACCORD trial group. Authors also acknowledge the authors of SSNHL study group to make their data publicly available and authors pay their sincere gratitude and regard to the members of the Doctoral advisory committee for their valuable suggestions.

## REFERENCES

[1] Harrell FE. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. New York: Springer; 2015.
[2] Collett D. Modelling survival data in medical research. CRC press. 3rd ed. 2015; Chapter-3: 57-73 & Chapter 8: 295-303.
[3] Rawlings JO, Pantula SG, Dickey DA, editors. Applied regression analysis: A research tool. New York, NY: Springer New York; 1998.
[4] Hickey GL, Kontopantelis E, Takkenberg JJ, Beyersdorf F. Statistical primer: Checking model assumptions with regression diagnostics. Interact Cardiovasc Thorac Surg. 2019;28(1):01-08.
[5] Arokiaraj SJ. Blood sugar, lipid profile and their correlation: A pilot study in Puducherry. Int J Appl Pure Sci Agric. 2016;2:140-44.
[6] Bønaa KH, Thelle DS. Association between blood pressure and serum lipids in a population. The tromsø study. Circulation. 1991;83(4):1305-14.
[7] Sharma S, Merchant J, Fleming SE. Lp (a)-cholesterol is associated with HDL-cholesterol in overweight and obese African American children and is not an independent risk factor for CVD. Cardiovascular Diabetology. 2012;11(1):01-07.
[8] Babler RL. Total blood cholesterol, blood triglyceride, and blood HDL correlation. Western Michigan University; 1991.
[9] Qabaha K, Hassan WA, Mansour H, Thanigachalam S, Naser S. Demographic and blood lipid profiles in correlation with heart attacks among mediterraneans. Journal of Nutrition & Food Sciences. 2014.1;4(4):01.
[10] Lee JS, Kim DH, Lee HJ, Kim HJ, Koo JW, Choi HG, et al. Lipid profiles and obesity as potential risk factors of sudden sensorineural hearing loss. PLoS One. 2015.10;10(4):e0122496.
[11] Mohammed AA. Lipid profile among patients with sudden sensorineural hearing loss. Indian J Otolaryngol Head Neck Surg. 2014;66(4):425-28.
[12] Chang IJ, Kang CJ, Yueh CY, Fang KH, Yeh RM, Tsai YT. The relationship between serum lipids and sudden sensorineural hearing loss: A systematic review and meta-analysis. PLoS One. 2015;10(4):e0121025.
[13] Tolonen N, Forsblom C, Mäkinen VP, Harjutsalo V, Gordin D, Feodoroff M, et al. Different lipid variables predict incident coronary artery disease in patients with type 1 diabetes with or without diabetic nephropathy: The FinnDiane study. Diabetes Care. 2014;37(8):2374-82.
[14] Sone H, Nakagami T, Nishimura R, Tajima N, MEGA Study Group. Comparison of lipid parameters to predict cardiovascular events in Japanese mild-to-moderate hypercholesterolemic patients with and without type 2 diabetes: Subanalysis of the MEGA study. Diabetes Res Clin Pract. 2016;113:14-22.
[15] Tohidi M, Hatami M, Hadaegh F, Safarkhani M, Harati H, Azizi F. Lipid measures for prediction of incident cardiovascular disease in diabetic and non-diabetic adults: Results of the 8.6 years follow-up of a population based cohort study. Lipids Health Dis. 2010;9(1):01-09.
[16] Sone H, Tanaka S, Tanaka S, Iimuro S, Ishibashi S, Oikawa S, et al. Comparison of various lipid variables as predictors of coronary heart disease in Japanese men and women with type 2 diabetes: Subanalysis of the Japan Diabetes Complications Study. Diabetes care. 2012;35(5):1150-57.
[17] Lee Y, Park S, Lee S, Kim Y, Kang MW, Cho S, et al. Lipid profiles and risk of major adverse cardiovascular events in CKD and diabetes: A nationwide population-based study. PloS one. 2020;15(4):e0231328.
[18] Shai I, Rimm EB, Hankinson SE, Curhan G, Manson JE, Rifai N, et al. Multivariate assessment of lipid parameters as predictors of coronary heart disease among postmenopausal women: Potential implications for clinical guidelines. Circulation. 2004;110(18):2824-30.
[19] Pedersen TR, Faergeman O, Kastelein JJ, Olsson AG, Tikkanen MJ, Holme I, et al. High-dose atorvastatin vs usual-dose simvastatin for secondary prevention after myocardial infarction: The IDEAL study: A randomized controlled trial. JAMA. 2005.16;294(19):2437-45.
[20] Ting ZWR, Yang X, Yu LW, Luk AO, Kong AP, Tong PC, et al. Lipid control and use of lipid-regulating drugs for prevention of cardiovascular events in Chinese type 2 diabetic patients: A prospective cohort study. Cardiovasc Diabetol. 2010;9(1):01-09.
[21] Paredes S, Fonseca L, Ribeiro L, Ramos H, Oliveira JC, Palma I. Novel and traditional lipid profiles in Metabolic Syndrome reveal a high atherogenicity. Scientific Reports. 2019;9(1):01-07.
[22] Hou Q, Li S, Gao Y, Tian H. Relations of lipid parameters, other variables with carotid intima-media thickness and plaque in the general Chinese adults: An observational study. Lipids Health Dis. 2018;17(1):107.
[23] Nissen SE, Tuzcu EM, Libby P, Thompson PD, Ghali M, Garza D, et al. Effect of antihypertensive agents on cardiovascular events in patients with coronary disease and normal blood pressure: The CAMELOT study: A randomized controlled trial. JAMA. 2004;292(18):2217-25.
[24] Xia W, Yao X, Chen Y, Lin J, Vielhauer V, Hu H. Elevated TG/HDL-C and non-HDL-C/HDL-C ratios predict mortality in peritoneal dialysis patients. BMC Nephrology. 2020;21(1):01-09.
[25] El-Fallah M, El-Salam A. A note on partial least squares regression for multicollinearity (A comparative study). International Journal of Applied Science and Technology. 2014;4(1):163-69.
[26] Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. PloS One. 2015;10(8):e0132382.
[27] Consultation WH. Obesity: Preventing and managing the global epidemic. World Health Organization technical report series. 2000;894:01-253.
[28] Johnson RA, Wichern DW. Applied multivariate statistical analysis. 5th ed. 2002; Chapter-6: 283-285.
[29] Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika. 1980;67(1):145-53.
[30] Mishra A, Harichandrakumar KT, Binu VS, Satheesh S, Nair NS. Multivariate approach in analyzing medical data with correlated multiple outcomes: An exploration using ACCORD trial data. Clinical Epidemiology and Global Health. 2021.1;11:100785.
[31] R Studio Team. RStudio: Integrated development for R. RStudio, Inc., Boston, MA URL http://www. rstudio. com. 2015;42(14):84.
[32] Statistics IS. IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp. Google Search. 2013.
[33] StataCorp LP. Stata multilevel mixed-effects reference manual. College Station, TX: StataCorp LP. 2013;9(10).

**PARTICULARS OF CONTRIBUTORS:**
1. PhD Scholar, Department of Biostatistics, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, Tamil Nadu, India.
2. Professor, Department of Biostatistics, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, Tamil Nadu, India.
3. Assistant Professor, Department of Biostatistics, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, Tamil Nadu, India.
4. Associate Professor, Department of Biostatistics, National Institute of Mental Health and Neurosciences, Bengaluru, Karnataka, India.
5. Professor, Department of Cardiology, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, Tamil Nadu, India.

**NAME, ADDRESS, E-MAIL ID OF THE CORRESPONDING AUTHOR:**
Dr. N Sreekumaran Nair,
Admin Block, 4th Floor, Department of Biostatistics, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, Tamil Nadu, India.
E-mail: nsknairmanipal@gmail.com

**PLAGIARISM CHECKING METHODS:** [Jain H et al.]
- Plagiarism X-checker: Feb 05, 2022
- Manual Googling: Apr 26, 2022
- iThenticate Software: Jun 02, 2022 (8%)

**ETYMOLOGY:** Author Origin

**AUTHOR DECLARATION:**
- Financial or Other Competing Interests: None
- Was Ethics Committee Approval obtained for this study? The IEC granted waiver of consent for the study.
- Was informed consent obtained from the subjects involved in the study? NA
- For any images presented appropriate consent has been obtained from the subjects. NA

Date of Submission: **Feb 02, 2022**
Date of Peer Review: **Mar 30, 2022**
Date of Acceptance: **Apr 27, 2022**
Date of Publishing: **Jul 01, 2022**