Development of Prediction Models under Multiple Imputation for Coronary Heart Disease in Type 2 Diabetes Mellitus

Guozhi Jiang[1], Eric S. Lau[1], Ying Wang[1], Andrea O. Luk[1], Claudia H. Tam[1], Janice S. Ho[1],
Vincent K. Lam[1], Heung M. Lee[1], Xiaodan Fan[2], Wing-Yee So[1], Juliana C. Chan[1]
and Ronald C. Ma[1]

[1]*Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, Hong Kong, SAR, China*
[2]*Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong, SAR, China*

Keywords: Multiple Imputation, Bootstrap, Prediction Model, Coronary Heart Disease, Type 2 Diabetes.

Abstract: The objectives of this study were to develop and compare the prediction models based on imputed data sets with that based on complete-case (C-C) data set for coronary heart disease (CHD) in type 2 diabetes mellitus (T2DM) and to identify novel genes associated with CHD from T2DM related genes. A prospective cohort of 5526 patients with T2DM and without known CHD and heart failure at baseline was used in this analysis. During a median follow-up time of 8.8 years, 406 (7.3%) patients developed incident CHD. Multiple imputation (MI) was performed to tackle missing values for 26 clinical variables and 40 genetic variables, while Cox proportional hazards regression with backward variable selection was applied to bootstrap samples. Five different MI or C-C models were compared and the performance based on C-index, 5 years AUC and the slope of prognostic index were similar, three SNPs located at NEGR1, CDKAL1 and ADAMTS9 were found to be significant after adjusting for clinical variables. In conclusion, multiple imputation and bootstrap can be benefit to the development of prediction model, and a stable risk factor set for CHD was successfully identified from our dataset containing clinical and genetic variables.

1 INTRODUCTION

The prevalence of type 2 diabetes mellitus (T2DM) is increasing around the world, and it leads to a 2-4 fold increased risk of coronary heart disease (CHD) compared to those patients without T2DM (Laakso, 2001). Based on Chinese diabetic population, (Yang et al., 2008) has developed a CHD prediction model using available clinical variables. Due to the complexity of CHD, however, this disease is influenced not only by lifestyle factors, but also by genetic factors (Vaarhorst et al., 2012). Simultaneously, taking account of the relationship between T2DM and CHD, it was hoped that T2DM associated genes were also associated with CHD and could be used to predict CHD risk.

In basis of practical problems in the application of model development, such as the processing of missing values and variable selection, we undertake this study to further investigate Yang's CHD prediction model. The objectives are to 1) perform multiple imputation (MI) to tackle the missing values and compare the performance of models developed from imputed data sets and from complete-case (C-C) data set for Yang's model, 2) select a stable CHD predictor list from a data set containing clinical and genetic variables, and 3) identify novel genes associated with CHD from T2DM related genes.

2 PATIENTS AND METHODS

2.1 Study Cohort

The data of this study was a cohort of the Hong Kong Diabetes Registry established in 1995 at a regional Hong Kong hospital. Among the total of 6013 unrelated T2DM patients (age 56.8±13.3 yr, 46% male) selected from this registry, 487 patients with known baseline CHD and heart failure were excluded. Therefore, a CHD prospective cohort including 5526 patients with detailed clinical information was used in this analysis.

CHD definition and detailed assessments ofmethods and laboratory assays were exactly the same as those described by (Yang et al., 2008). To

Jiang G., S. Lau E., Wang Y., O. Luk A., H. Tam C., S. Ho J., K. Lam V., M. Lee H., Fan X., So W., C. Chan J. and C. Ma R..
Development of Prediction Models under Multiple Imputation for Coronary Heart Disease in Type 2 Diabetes Mellitus.
DOI: 10.5220/0004244303120315
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2013), pages 312-315
ISBN: 978-989-8565-35-8

Table 1: Parameter estimates for two CHD clinical models in Chinese type 2 diabetic patients.

Variable	MI Clinical Model				C-C Clinical Model			
	Beta	SE	HR (95% CI)	P	Beta	SE	HR (95% CI)	P
Age	0.035	0.005	1.04 (1.03-1.04)	<0.001	0.034	0.005	1.03 (1.02-1.04)	<0.001
Male sex	0.334	0.108	1.4 (1.13-1.72)	0.002	0.301	0.113	1.35 (1.08-1.69)	0.008
Current smoker	0.364	0.143	1.44 (1.09-1.91)	0.011	0.348	0.150	1.42 (1.06-1.9)	0.020
Duration of diabetes	0.029	0.007	1.03 (1.02-1.04)	<0.001	0.029	0.007	1.03 (1.01-1.04)	<0.001
Log10(ACR)	0.277	0.078	1.32 (1.13-1.54)	<0.001	0.272	0.080	1.31 (1.12-1.54)	<0.001
Log10(eGFR)	-1.182	0.306	0.31 (0.17-0.56)	<0.001	-1.254	0.344	0.29 (0.15-0.56)	<0.001
Non-HDL cholesterol	0.208	0.043	1.23 (1.13-1.34)	<0.001	0.214	0.045	1.24 (1.13-1.35)	<0.001

compare with Yang's model, the same list of 26 baseline clinical variables as Yang used was included in this study. As the total cholesterol was linear with non-high-density lipoprotein cholesterol and high-density lipoprotein cholesterol, and the family history of CHD was not available in this dataset, they were excluded from this candidate list. In addition, 40 published single-nucleotide polymorphisms (SNPs) known to be associated with T2DM in genome-wide association study were genotyped. All SNPs passed quality control for Hardy-Weinberg equilibrium, minor allele frequency and SNP call rate.

2.2 Multiple Imputation and Variable Selection by Bootstrap

To impute the missing values, Multiple Imputation via Chained Equations (MICE) procedure was performed according to the guidelines described by (van Buuren et al., 1999). In short, predictive mean matching and polytomous regression were specified for continuous and categorical variables, respectively. All 7 clinical variables selected in Yang's final model, as well as the outcome variable and the natural logarithm of survival time, were always kept. As it was recommended that the suitable number of variables used in each imputation model should be no more than 25, we set the cut-off value of correlation to 0.1 for clinical variables and 0.03 for genetic variables. After those steps, a series of imputation models that consisted of the best 10 to 23 predictor variables were built. Usually 5 to 10 repeated imputations would be enough to achieve high efficiency, here we generated 10 imputed data sets. Moreover, Rubin's rules were used to combine the regression coefficients and variances.

To take the sampling variation into account and get a stable variable subset, we applied Cox proportional hazards regression with backward variable selection (p <0.05 for stay) to 100 bootstrap samples for each of 10 imputed data sets. We calculated the inclusion frequency for each variable appearing in 1000 variable subsets, and selected the top variables with frequencies more than 50% to develop models on each imputed data set. Finally, those models were pooled into a final model by Rubin's rules.

2.3 Development of Prediction Models

To compare the performance of different models based on imputed data sets and C-C data set, two pairs of MI models and C-C models were developed: 1) MI Clinical Model and C-C Clinical Model. Both models were developed using 7 clinical variables that Yang selected. 2) MI Final Model and C-C Final Model. These two models were developed using the top clinical and genetic variables selected by variable selection on imputed data sets. Furthermore, in order to measure the model variation induced by training/test split method as Yang used, we also constructed the C-C Split Models based on 7 clinical variables. We randomly split the C-C data into 1:1 training and test sets for 100 times, and training sets were used to develop models while test sets were used to evaluate performance.

The performance of the MI models or C-C models was measured by discrimination and calibration. Three different evaluation methods were employed to measure the discrimination: overall C index, time-dependent area under the curve (AUC) of receiver operator characteristics (ROC) curves with 5 years specified and category-free net reclassification improvement (NRI). Meanwhile, we used the slope of the prognostic index (PI) to quantify calibration. Furthermore, bootstrapping method as described by (Harrel Jr, 1996) was applied to provide nearly unbiased estimates of predictive performance.

Table 2: Parameter estimates for two CHD final models in Chinese type 2 diabetic patients.

Variable	Inclusion Frequency	MI Final Model				C-C Final Model			
		Beta	SE	HR (95% CI)	P	Beta	SE	HR (95% CI)	P
Age	99.8%	0.029	0.005	1.03 (1.02-1.04)	<0.001	0.027	0.006	1.03 (1.02-1.04)	<0.001
Duration of diabetes	95.6%	0.030	0.007	1.03 (1.02-1.04)	<0.001	0.029	0.008	1.03 (1.01-1.04)	<0.001
Log10(eGFR)	89.8%	-1.111	0.310	0.33 (0.18-0.60)	<0.001	-1.298	0.367	0.27 (0.13-0.56)	<0.001
HDL cholesterol	87.5%	-0.566	0.159	0.57 (0.42-0.78)	<0.001	-0.575	0.182	0.56 (0.39-0.8)	0.002
Peripheral arterial disease	78.4%	0.458	0.154	1.58 (1.17-2.14)	0.003	0.494	0.170	1.64 (1.18-2.29)	0.004
rs2568958	72.8%	0.313	0.122	1.37 (1.08-1.74)	0.010	0.373	0.130	1.45 (1.13-1.87)	0.004
rs7754840	69.3%	-0.211	0.074	0.81 (0.70-0.94)	0.004	-0.191	0.083	0.83 (0.7-0.97)	0.021
Log10(ACR)	64.2%	0.233	0.082	1.26 (1.08-1.48)	0.004	0.264	0.091	1.3 (1.09-1.56)	0.004
LDL cholesterol	60.1%	0.213	0.048	1.24 (1.13-1.36)	<0.001	0.232	0.055	1.26 (1.13-1.41)	<0.001
Male sex	53.3%	0.288	0.109	1.33 (1.08-1.65)	0.008	0.306	0.124	1.36 (1.07-1.73)	0.013
Current smoker	51.1%	0.348	0.144	1.42 (1.07-1.88)	0.016	0.405	0.162	1.5 (1.09-2.06)	0.012
Systolic BP	50.2%	0.005	0.003	1.01 (1-1.01)	0.037	0.004	0.003	1 (1-1.01)	0.151
rs4607103	50.1%	0.150	0.076	1.16 (1-1.35)	0.049	0.141	0.084	1.15 (0.98-1.36)	0.094

3 RESULTS

3.1 Cohort Description

Of the total 5526 T2DM patients in CHD prospective cohort, 406 (7.3%) were found to develop CHD during a median follow-up period of 8.8 (IQR: 6.0-11.4) years. Patients who progressed to CHD were significantly older, higher BP, higher HbA1c, had a longer duration of diabetes, and were more likely to use drugs, compared to those who didn't develop CHD. For the missing value percentage of each clinical and genetic variable, most of them were less than 10%, only one variable (rs10838738) reached 15% while 16 variables had no missing values.

3.2 Performance of Prediction Models

The estimates of parameters for MI Clinical Model and C-C Clinical Model were similar (Table 1). All 7 factors were significant and the effects were close in both models, but the MI Clinical Model had a relative lower standard error for each factor. The biased-corrected C-index and 5-years AUC were very close, and both models showed good calibration (Table 3). When comparing with C-C Split Models, the performance was also similar (0.734 vs. 0.728 for C-index, 0.738 vs. 0.732 for AUC), but the ranges of indicators were larger in C-C Split Models. The C-index and AUC for Yang's CHD model were 0.704 and 0.737 respectively; these values were also included in this range.

Table 2 presents the MI Final Model and C-C Final Model, all selected factors were significant in imputed data sets, but systolic BP and rs4607106 were not significant in C-C data set. The effect of each factor was similar, and standard error was lower in MI Final Model. When comparing with Yang's model, the selected factor list was a little different. 6 variables were included in both models, but our model selected HDL cholesterol and LDL cholesterol instead of non-HDL cholesterol, and 2 other variables (peripheral arterial disease and systolic BP) as well as 3 SNPs (rs2568958, rs7754840 and rs4607103). From table 3, the performance of MI Final Model and C-C Final Model was close, but the MI Final Model was slightly better than MI Clinical Model (0.744 vs. 0.734 for C-index, and 0.748 vs. 0.738 for AUC). When considering the impact of SNPs to prediction model, the NRI was 14.6% for MI Final Model but only 2.5% for C-C Final Model.

Table 3: Bias-corrected predictive performance for five different models.

	MI Final Model*	MI Clinical Model*	C-C Final Model	C-C Clinical Model	C-C Split Models*
C-index	0.744 [0.742-0.744]	0.734 [0.733-0.738]	0.747	0.731	0.728 [0.683-0.753]
AUC	0.748 [0.747-0.749]	0.738 [0.736-0.741]	0.749	0.732	0.732 [0.678-0.765]
Slope	0.961 [0.954-0.966]	0.981 [0.976-0.992]	0.949	0.975	0.956 [0.67-1.209]
NRI	14.6% [10%-17.9%]	/	2.47%	/	/

* Data are expressed as median [full range].

4 DISCUSSION

In this study, we have given an example to illustrate the process of prediction model development based on incomplete data. To get a more stable risk factor set from clinical and genetic variable list for CHD in T2DM, we integrated bootstrap and backward variable selection on imputed data sets.

Incomplete data are commonly encountered in medical research. Excluding all patients with any missing values may lose useful information and reduce the power of prediction model, which leads to some variables not attaining statistical significance, such as for the systolic BP and rs4607106 in our MI Final Model and C-C Final Model. In our study, the MI models are very similar to the C-C models, it is because the missing rates are not high and the sample sizes are close, but imputation makes it more powerful to perform variable selection.

Combining bootstrap resampling with variable selection will be benefit to the stability of selected variables. Through bootstrap and variable selection, variables with strong effects on the outcome will be selected more frequently than those with no or weak effects. To validate a model, data-splitting as a simple method is commonly used, but the model performance will vary greatly with different splits, and bias will be introduced. Our results showed the bootstrapping bias-corrected indicators of performance were close to the median indicators produced by multiple times training/test splits. Therefore, to ensure an honest model evaluation, we would better evaluate the models by generating multiple pairs of training/test sets or use bias-corrected method.

Importantly, three SNPs (rs2568958, rs7754840 and rs4607103 located at NEGR1, CDKAL1 and ADAMTS9 gene, respectively) were selected with high inclusion frequencies and the NRI results indicated they contributed to the CHD prediction. Therefore, these three T2D-related SNPs may also have association effects with CHD. To validate the effect of these SNPs, we will try to do some further analyses, such as replication study.

In conclusion, this cohort study illustrated the MICE and bootstrap can be benefit to the development of prediction model based on dataset containing clinical and genetic variables. An informative risk factor set for CHD, including three T2D-related SNPs, was successfully identified from CHD prospective cohort of Hong Kong Chinese patients with T2DM. Future research will be needed to validate the effect of these selected SNPs.

ACKNOWLEDGEMENTS

This work was supported by the Innovation and Technology Fund (ITS/487/09FP), RGC Central Allocation Scheme (CUHK 1/04C), RGC Earmarked Research Grant (CUHK4724/07M), and the CUHK Direct Grant (2150476 and 2141611).

REFERENCES

Harrel Jr, F. E. a. L., K. L. and Mark, D. B. 1996. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error. *Statistics in Medicine***,** 361–387.

Laakso, M. 2001. Cardiovascular disease in type 2 diabetes: challenge for treatment and prevention. *J Intern Med,* 249**,** 225-35.

Vaarhorst, A. A., Lu, Y., Heijmans, B. T., Dolle, M. E., Bohringer, S., Putter, H., Imholz, S., Merry, A. H., van Greevenbroek, M. M., Jukema, J. W., Gorgels, A. P., van den Brandt, P. A., Muller, M., Schouten, L. J., Feskens, E. J., Boer, J. M. & Slagboom, P. E. 2012. Literature-based genetic risk scores for coronary heart disease: the Cardiovascular Registry Maastricht (CAREMA) prospective cohort study. *Circ Cardiovasc Genet,* 5**,** 202-9.

van Buuren, S., Boshuizen, H. C. & Knook, D. L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med,* 18**,** 681-94.

Yang, X., So, W. Y., Kong, A. P., Ma, R. C., Ko, G. T., Ho, C. S., Lam, C. W., Cockram, C. S., Chan, J. C. & Tong, P. C. 2008. Development and validation of a total coronary heart disease risk score in type 2 diabetes mellitus. *Am J Cardiol,* 101**,** 596-601.