Bootstrap Variance Estimation for Predicted Individual and Population-Average Risks

Milorad S. Kovacevic, Lenka Mach, Georgia Roberts
Statistics Canada, Ottawa, ON, Canada, K1A0T6

Abstract
Often there is a need to predict the probability that an individual with specific characteristics (risk factors) will suffer from a certain disease or a health condition. Sometimes the real interest is in providing the aggregate predicted risk estimates at the population level or at the level of a subpopulation (e.g., males aged 65+). A related problem is prediction of such a probability for an individual who does not belong to the surveyed population. While the estimation of these probabilities follows from fitting an appropriate model to the available data, the design-based standard error estimation of so-obtained estimates is not obvious. We are proposing a bootstrap method for estimation of the standard errors of predicted individual and aggregate risks. The method is illustrated using data from the Canadian Community Health Survey.

Key Words: Logistic model, prediction models, standard error, validation sample, survey bootstrap, design-based variance

1. Introduction

The fitting of statistical models to predict risks of having or developing a disease or a health condition usually has two objectives: (i) Predict an individual's risk of disease during a specified time range given his specific risk factors. (ii) Predict the population average risk of disease during a specified time range assuming a certain risk-factor distribution in the population. The former is known as a conditional prediction, the conditioning being on the given set of risk - factor values, while the latter is known as a marginal prediction, with the 'marginalization' over the assumed risk-factor distribution in the population.

Prediction models are frequently used by medical researchers to predict individual risks. Some good examples are the Framingham Risk Models developed to predict the risk of cardio-vascular diseases (for a recent reference see D'Agostino RB, et al., 2008). Some of these models are based on samples taken from the general population using complex sample designs.

There are several sources of uncertainty when predicting risks: the incomplete knowledge of the determinants and risk factors of health that govern the prevalence or incidence of diseases over time, the inherent stochastic nature of events and their order, a sample-based inference about the model and population, the presence of measurement errors, etc.

Logistic models as well as survival models are typically used to estimate the risks of having or developing a disease for the members of the population by individual demographic, lifestyle and risk profiles. The models are developed using the observed association of these profiles with disease occurrence in survey data, possibly linked to administrative health databases. A good example is the study on diabetes mellitus prediction based on the Ontario portion of the Canadian National Population Health Survey conducted by Statistics Canada, which was linked to the Ontario Diabetes Database, a population-based registry (Rosella, Manuel, Burchill, Stukel, 2008). These models are then used to predict individual and population average risks.

Our objective in this paper is to estimate the variability of the prediction, either conditional or marginal, generated by randomness of the sample used for estimation of the prediction model, that is, to estimate $E(\hat{p}-E(\hat{p}))^2$, where $\hat{p}$ is the estimated predicted value, and the expectation is taken over the sampling-induced randomness.

The validation of a model, as well as the estimation of the prediction error, have to account for the complexity of the survey design. In this paper we are interested just in the latter; that is, we assume that the model validation was done properly, and that the model is acceptable for making predictions.

We propose the estimation of the standard errors for predicted risks by the survey bootstrap method (Rao, Wu, Yue, 1992). If the model estimated from one sample is to be used for estimating risks with another sample, it is assumed that bootstrap weights are available for both the sample used for fitting the prediction model (a training sample) and for the sample from which we predict the population average risks.

The paper is organized as follows. In the next section we briefly review the logistic model that is frequently used to predict the risks of developing a disease. Section 3 deals with the estimation of the standard error of the prediction by the survey bootstrap method. In Section 4 we address the situation where the model is developed using one sample but the prediction is based on another sample. We illustrate the method in Section 5 using data from the Canadian Community Health Survey (CCHS). Section 6 contains an overall summary. Throughout the paper, the “probability” of having or developing a disease and the “risk” will be used interchangeably.

2. Prediction of Having or Developing a Disease

Let us assume that the probability of having or developing a disease is modeled as a function (non-linear) of given *x* and β, and that it varies from individual to individual, where, for individual *i*:

$$Prob\left(y_i = 1 \middle| x_i, \beta\right) = p(x_i, \beta).$$

Here β and x_i represent vectors of unknown model parameters and values of risk factors for individual *i*, respectively. In the case of the logistic regression model, this probability is given by

$$p(x_i, \beta) = \frac{1}{1 + e^{-x_i'\beta}} \ . \tag{1}$$

The predicted probability for a person with the health and lifestyle profile including the risk factors denoted by x_0 is

$$\hat{p}_0\left(\hat{\beta}\right) = \hat{p}\left(x_0, \hat{\beta}\right) = \frac{1}{1 + e^{-x_0'\hat{\beta}}} \ , \tag{2}$$

where $\hat{\beta}$ is the vector of regression coefficients estimated from a sample representing the population to which the person belongs. In this way we obtain the conditional prediction of the probability for the person with the profile x_0. The variability of the individual prediction is determined only by the variability of $\hat{\beta}$. Note that $x_0'\ \hat{\beta}\ (= \hat{\eta}_0)$ is usually referred to as the estimated risk score or prognostic index for this person.

The average (marginal) probability for the population is predicted by the weighted mean of the individual (conditional) predictions:

$$\hat{p}\left(\hat{\beta}\right) = \sum_{i \in s} w_i^* \hat{p}_i\left(\hat{\beta}\right), \tag{3}$$

where $\hat{p}_i\left(\hat{\beta}\right) = p\left(x_i, \hat{\beta}\right)$, and the weights $w_i^* = w_i \Big/ \sum_{j \in s} w_j$ are the survey weights for sample *s* scaled to sum up to 1. The variability of $\hat{p}\left(\hat{\beta}\right)$ is determined by the sample design and resulting survey weights <u>and</u> by the variability of $\hat{\beta}$.

The total count of individuals having or developing the disease for the population can be predicted as $\hat{p}\left(\hat{\beta}\right)\hat{N}$ where $\hat{p}\left(\hat{\beta}\right)$ is given by (3) and $\hat{N}$ represents the predicted size of the population.

If the marginal probability is being predicted just for a subpopulation of the population covered by the survey, s in (3) denotes the sample that falls in that subpopulation, rather than the full sample.

3. Estimation of Variance of Predictions by Bootstrapping

Estimating the variability of predictions is very important where acting upon the predictions may have medical or financial consequences. Estimated standard errors (which are the square root of the estimated variances) are usually used to obtain confidence intervals for the predictions.

In this section we assume that the prediction model is correctly specified. Also, we assume that the sample data are obtained by a complex (multi-stage) sample design, and that the parameters of the model are estimated properly accounting for the sample design, *i.e.*, that suitable survey weighted estimates of these parameters are obtained.

Further, we assume that the bootstrapping of primary sampling units (PSU's) with replacement (see Rao-Wu (1988), as specified in Rao, Wu, Yue (1992)), is an appropriate method for design-based variance estimation, and that a large number, say B, of sets of bootstrap weights, $w_i^{(b)}$, b=1,2,…,B, is available.

3.1 Bootstrap Variance Estimation for the Estimated β Coefficients

Let $\hat{\beta}$ be a full-sample estimate of the coefficient (vector) β in the model discussed in Section 2. Let $\hat{\beta}^{(b)}$ be the estimate of β obtained when using the bth set of bootstrap weights (b=1,2,…, B).

The bootstrap estimate of the variance-covariance matrix of $\hat{\beta}$ is

$$\hat{V}\left(\hat{\beta}\right)= \sum_{b=1}^{B}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)\left(\hat{\beta}^{(b)} - \hat{\beta}\right)' \Big/ B \ .$$

Binder, Kovacevic and Roberts (2005) studied the performance of several different bootstrap estimators for variance estimation of $\hat{\beta}$ by means of a simulation study and found generally good performance of the above estimator regarding the bias and the efficiency. The variance of the individual risk score, $\hat{\eta}_0 = x_0'\hat{\beta}$, can be estimated as $\hat{V}(\hat{\eta}_0) = x_0' \, \hat{V}\left(\hat{\beta}\right)x_0$ where $\hat{V}\left(\hat{\beta}\right)$ is the bootstrap estimate defined above.

3.2 Bootstrap Variance Estimation for the Conditional (Individual) Prediction

First we assume that the person for whom the probability of "the event" is predicted belongs to the population represented by the sample used to fit the model. The predicted probability is estimated by (2). Using the B sets of bootstrap weights it is also possible to calculate bootstrap replicates of the predicted probability for this individual:

$$\hat{p}_0\left(\hat{\beta}^{(b)}\right)= p\left(x_0, \hat{\beta}^{(b)}\right),\ b=1,\ldots,B.$$

The bootstrap variance estimate of $\hat{p}_0\left(\hat{\beta}\right)$ is

$$\hat{V}\left(\hat{p}_0\left(\hat{\beta}\right)\right)= \sum_{b=1}^{B}\left[\hat{p}_0\left(\hat{\beta}^{(\mathrm{b})}\right)- \hat{p}_0\left(\hat{\beta}\right)\right]^2 \Big/ B \ . \tag{4}$$

Estimator (4) captures the variability of the conditional prediction that is generated by the variability of the estimated β. The statistical properties of this estimator are yet to be investigated.

3.3 Confidence Limits for the Conditional Predicted Probability

Although $\hat{\beta}$ is approximately normally distributed, it is not advisable to assume a normal distribution for the predicted probability (which is a non-linear function of $\hat{\beta}$), and thus, it is not recommended to use a symmetric normal approximation of the prediction limits, $\hat{p}_0\left(\hat{\beta}\right)\mp z_{\alpha/2}\sqrt{\hat{V}\left(\hat{p}_0\left(\hat{\beta}\right)\right)}$. Also, for small probabilities this interval may have a negative lower bound. Instead, it is better to use an approximation based on a transformation of the normal prediction limits for the risk score $\hat{\eta}$ which is linear in $\hat{\beta}$, and thus closer to normality than $\hat{p}_0\left(\hat{\beta}\right)$. A confidence interval obtained by this approach is

$$\left\{\left[1+\boldsymbol{exp}\left(-\hat{\eta}_0+z_{\alpha/2}\sqrt{\hat{V}(\hat{\eta}_0)}\right)\right]^{-1}\boldsymbol{,}\left[1+\boldsymbol{exp}\left(-\hat{\eta}_0-z_{\alpha/2}\sqrt{\hat{V}(\hat{\eta}_0)}\right)\right]^{-1}\right\}.$$

It is asymmetric and usually longer than the normal approximation, and is expected to have better coverage properties under the assumed risk model. For a discussion on transformation of confidence intervals for the proportions see Rust and Rao (1996).

Alternatively, if a sufficiently large number of bootstrap weights is available, one can use the bootstrap prediction limits obtained as percentiles of the distribution of bootstrap replicates of the predicted individual probabilities $\hat{p}_0\left(\hat{\beta}^{((b)}\right),\ b=\mathbf{1,...,}B$. Such an interval takes the form: $\left\{p_{[(B+1)\alpha/2]},\quad p_{[(B+1)(1-\alpha/2)]}\right\}$, with the limits representing the $[(B+1)\alpha/2]$-th and the $[(B+1)(1-\alpha/2)]$-th ordered values of B replicate values $\hat{p}_0\left(\hat{\beta}^{((b)}\right)$ *i.e.*, the bootstrap percentiles (Shao and Tu, 1995).

3.4 Bootstrap Variance Estimation for the Marginal (Population Average) Prediction

The population average probability of "the event" is predicted by (3), *i.e.*, by the weighted mean of the predicted individual probabilities:

$$\hat{p}\left(\hat{\beta}\right)=\sum_{i\in s}w_i^*\hat{p}_i\left(\hat{\beta}\right). \tag{5}$$

Here, the sampling variability enters the estimating equation in two different ways - both through the estimation of the beta parameters and also through the prediction of the average (marginal) probability as a survey-weighted mean of predicted individual probabilities. The survey bootstrap method will take into account both of these variabilities. Thus the bootstrap variance estimate for the marginal (population average) prediction is given by

$$\hat{V}\ \left(\hat{p}\left(\hat{\beta}\right)\right)=\frac{1}{B}\sum_{b=1}^{B}\left[\hat{p}^{(b)}\left(\hat{\beta}^{\boldsymbol{(}b\boldsymbol{)}}\right)-\hat{p}\left(\hat{\beta}\right)\right]^2 \tag{6}$$

where $\hat{p}^{(b)}\left(\hat{\beta}^{\boldsymbol{(}b\boldsymbol{)}}\right)=\sum_{i\in s}w_i^{*(b)}\hat{p}_i\left(\hat{\beta}^{\boldsymbol{(}b\boldsymbol{)}}\right)$ is obtained using the *b*-th set of bootstrap weights.

Confidence limits for the population average probability can be the normal approximation prediction limits $\hat{p}\left(\hat{\beta}\right)\mp z_{\alpha/2}\sqrt{\hat{V}\left(\hat{p}\left(\hat{\beta}\right)\right)}$ if $\hat{p}\left(\hat{\beta}\right)$ can be assumed to be approximately normally distributed. Otherwise, it would be advisable to find a transformation of $\hat{p}\left(\hat{\beta}\right)$ that could be assumed to be normal.

Note that if β is assumed to be known constant, an estimate of the variance of $\hat{p}(\hat{\beta})$ reduces to

$$\hat{V}_1\left(\sum_{i\in s}w_i^*\hat{p}_i\left(\hat{\beta}\right)\middle|\ \hat{\beta}=\beta\right)=\frac{1}{B}\sum_{d=1}^{B}\left[\hat{p}^{\boldsymbol{(}d\boldsymbol{)}}(\beta)-\hat{p}(\beta)\right]^2. \tag{7}$$

Such an assumption is sometimes made when estimating the variance of a complex statistic dependant on an unknown parameter assumed to be nuisance. The difference between (6) and (7) can be interpreted as the portion of the variance of the marginal prediction attributive to the variability of $\hat{\beta}$**.**

4. Predicting the Population Average Risk From a Different Sample

Suppose that we would like to apply the model estimated from one sample (a "training sample" s_t) to individuals from another independent sample, s_v, in order to estimate the population average risk of having or developing the disease.

Two situations where this might be done are the following:

(i) The sample s_v is taken from the same population as s_t and is being used for validation of the prediction model. For the rest of this section we will call s_v the validation sample.

(ii) The prediction model estimated from s_t is applied to a sample s_v taken from another population, where it is assumed that there is no difference in the populations from which these samples were taken. For example, a model is estimated from a sample from one province (or state) and then applied to a sample from another province where it is assumed that the relation between risk factors and the development of the disease is the same in both provinces, *i.e.*, the same underlying prediction model is assumed to hold in both.

The two samples are not necessarily obtained using the same sample design. We also assume that there are *B* sets of bootstrap replicate weights available for the s_t sample and *D* sets of replicate weights for the s_v sample.

From the training sample, s_t, and its bootstrap replicates, we obtain $\hat{\beta}_t$, as well as $\hat{\beta}_t^{(b)}, b = 1,\ldots,B$. Subscript *t* refers to the "training sample" and it means that the weights from the training sample were used.

Using $\hat{\beta}_t$ and $\hat{\beta}_t^{(b)}$ together with the observed risk factors x_v from validation sample s_v, we obtain individual predictions and their bootstrap replicates for $i \in s_v$:

$$\hat{p}_{v,i}\left(\hat{\beta}_t\right) = p\left(x_{v,i}, \hat{\beta}_t\right) \text{ and } \hat{p}_{v,i}\left(\hat{\beta}_t^{(b)}\right) = p\left(x_{v,i}, \hat{\beta}_t^{(b)}\right).$$

A prediction of the average probability for the population represented by the validation sample s_v is then obtained as a mean of the individual predictions,

$$\hat{p}_v\left(\hat{\beta}_t\right) = \sum_{i \in s_v} w_{v,i}^* \, \hat{p}_{v,i}\left(\hat{\beta}_t\right) , \qquad (8)$$

where $w_{v,i}^* = w_{v,i} \Big/ \sum_{i \in s_v} w_{v,i}$ are the survey weights of the validation sample s_v scaled to sum to 1, and *D* bootstrap replicates of the average risk prediction are

$$\hat{p}_v^{(d)}\left(\hat{\beta}_t\right) = \sum_{i \in s_v} w_{v,i}^{*(d)} \, \hat{p}_{v,i}\left(\hat{\beta}_t\right), \quad d = 1,\ldots, D .$$

The variance of $\hat{p}_v(\hat{\beta}_t)$, assuming that samples s_v and s_t are independent is given by:

$$V\left(\hat{p}_v\left(\hat{\beta}_t\right)\right) = E_1 \; V_2\left(\sum_{i \in s_v} w_{v,i}^* \hat{p}_{v,i}\left(\hat{\beta}_t\right) \middle| \; \hat{\beta}_t\right) + V_1 \, E_2\left(\sum_{i \in s_v} w_{v,i}^* \hat{p}_{v,i}\left(\hat{\beta}_t\right) \middle| \hat{\beta}_t\right) . \qquad (9)$$

The E_1 and V_1 in (9) are estimated using the training sample and its bootstrap replicates, while E_2 and V_2 are estimated from the s_v sample and its bootstrap replicates.

In the first term on the right-hand side (rhs) of (9), the variance $V_2\left(\sum_{i \in s_v} w_{v,i}^* \hat{p}_{v,i}\left(\hat{\beta}_t\right) \middle| \; \hat{\beta}_t\right)$ accounts for the sampling variability of the predicted mean assuming that $\hat{\beta}_t$ is fixed (non-random). This variance can be estimated by the bootstrap variance estimator

$$\hat{V}_2\left(\hat{p}_v\left(\hat{\beta}_t\right) \middle| \; \hat{\beta}_t\right) = \frac{1}{D} \sum_{d=1}^{D} \left[\hat{p}_v^{(d)}\left(\hat{\beta}_t\right) - \hat{p}_v\left(\hat{\beta}_t\right)\right]^2 . \qquad (10)$$

The expectation E_1 in (9) is over the sampling distribution of $\hat{\beta}_t$ and can be estimated by the distribution of the bootstrap replicates of the predicted probabilities, *i.e.*, the expectation E_1 can be estimated by a mean over the bootstrap replicates of the predicted probabilities. Thus, $E_1 V_2$ is estimated by
:

$$\hat{E}_1 \hat{V}_2 \left[\hat{p}_v \left(\hat{\beta}_t \right) \middle| \hat{\beta}_t \right] = \frac{1}{B} \sum_{b=1}^{B} \hat{V}_2 \left(\hat{p}_v \left(\hat{\beta}_t^{(b)} \right) \middle| \hat{\beta}_t^{(b)} \right) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{D} \sum_{d=1}^{D} \left[\hat{p}_v^{(d)} \left(\hat{\beta}_t^{(b)} \right) - \hat{p}_v \left(\hat{\beta}_t^{(b)} \right) \right]^2 \qquad (11)$$

Alternatively, $E_1 V_2$ in (9) can be estimated by (10), *i.e.*, by $\hat{V}_2$ itself.

In the second term on the rhs of (9) the expectation E_2 is taken over the sampling distribution of the population-average probability assuming that $\hat{\beta}_t$ is fixed (non-random). Expectation E_2 can be estimated by the mean of the bootstrap replicates of the population average probability $\hat{p}_v^{(d)} \left(\hat{\beta}_t \right) = \sum_{i \in s_v} w_{v,i}^{*(d)} \hat{p}_{v,i} \left(\hat{\beta}_t \right)$:

$$\hat{E}_2 \left(\sum_{i \in s_v} w_{v,i}^{*} \hat{p}_{v,i} \left(\hat{\beta}_t \right) \middle| \hat{\beta}_t \right) = \frac{1}{D} \sum_{d=1}^{D} \sum_{i \in s_v} w_{v,i}^{*(d)} \hat{p}_{v,i} \left(\hat{\beta}_t \right) = \hat{\bar{p}}_v \left(\hat{\beta}_t \right). \qquad (12)$$

Alternatively, E_2 can be estimated by the population average probability itself:

$$\hat{E}_2 \left(\sum_{i \in s_v} w_{v,i}^{*} \hat{p}_{v,i} \left(\hat{\beta}_t \right) \middle| \hat{\beta}_t \right) = \sum_{i \in s_v} w_{v,i}^{*} \hat{p}_{v,i} \left(\hat{\beta}_t \right) = \hat{p}_v \left(\hat{\beta}_t \right). \qquad (13)$$

The variance V_1, taken over the distribution of predicted individual probabilities, is estimated using the distribution of the bootstrap replicates of the predicted probabilities. Thus the variance $V_1(\hat{E}_2)$ is estimated by the following bootstrap variance when $\hat{E}_2$ is given by (12):

$$\hat{V}_1 \hat{E}_2 \left(\sum_{i \in s_v} w_{v,i}^{*} \hat{p}_{v,i} \left(\hat{\beta}_t \right) \middle| \hat{\beta}_t \right) = \hat{V}_1 \left(\hat{\bar{p}}_v \left(\hat{\beta}_t \right) \right) = \frac{1}{B} \sum_{b=1}^{B} \left[\hat{\bar{p}}_v \left(\hat{\beta}_t^{(b)} \right) - \hat{\bar{p}}_v \left(\hat{\beta}_t \right) \right]^2 . \qquad (14)$$

Alternatively, when $\hat{E}_2$ is given by (13),

$$\hat{V}_1 \hat{E}_2 \left(\sum_{i \in s_v} w_{v,i}^{*} \hat{p}_{v,i} \left(\hat{\beta}_t \right) \middle| \hat{\beta}_t \right) = \hat{V}_1 \left(\hat{p}_v \left(\hat{\beta}_t \right) \right) = \frac{1}{B} \sum_{b=1}^{B} \left[\hat{p}_v \left(\hat{\beta}_t^{(b)} \right) - \hat{p}_v \left(\hat{\beta}_t \right) \right]^2 . \qquad (15)$$

Finally, the variance of the population average probability (8) is approximated by the sum of the estimated components, and takes the following forms

$$\hat{V}_{A1} \left(\hat{p}_v \left(\hat{\beta}_t \right) \right) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{D} \sum_{d=1}^{D} \left[\hat{p}_v^{(d)} \left(\hat{\beta}_t^{(b)} \right) - \hat{p}_v \left(\hat{\beta}_t^{(b)} \right) \right]^2 + \frac{1}{B} \sum_{b=1}^{B} \left[\hat{\bar{p}}_v \left(\hat{\beta}_t^{(b)} \right) - \hat{\bar{p}}_v \left(\hat{\beta}_t \right) \right]^2 , \qquad (16)$$

$$\hat{V}_{A2} \left(\hat{p}_v \left(\hat{\beta}_t \right) \right) = \frac{1}{D} \sum_{d=1}^{D} \left[\hat{p}_v^{(d)} \left(\hat{\beta}_t \right) - \hat{p}_v \left(\hat{\beta}_t \right) \right]^2 + \frac{1}{B} \sum_{b=1}^{B} \left[\hat{p}_v \left(\hat{\beta}_t^{(b)} \right) - \hat{p}_v \left(\hat{\beta}_t \right) \right]^2 . \qquad (17)$$

5. Illustration

Note that this example is made to illustrate only the methods presented in this paper and should not be interpreted in a subject-matter context.

We use data from the Canadian Community Health Survey (CCHS) Cycle 3.1 to illustrate the method for variance estimation of the predicted probabilities of having or developing a disease. Only the sample from the province of Ontario (37,430 respondents) was used to estimate the prediction model, *i.e.*, to estimate the beta parameters. The dependent variable in the model is the binary variable indicating whether the respondent has coronary heart disease (CHD). The independent variables (risk factors) considered in the model are represented by five binary variables: Sex (SEX), being a former daily smoker (DAILY SMK), having more than high school education (HIGH SCH+), having diabetes (DIABETES), and having high blood pressure (HIGH BP); and by two continuous variables: AGE (in years) and body mass index (self reported) (BMI). Note that we also used AGE^2 in the model. It was marginally significant. Table 1 contains estimates of the beta parameters and their standard errors obtained from 500 bootstrap replicates. All the variables were significant at the 5% level. The estimates in Table 1 were produced using SUDAAN.

Table 1. Estimated Parameters of the Prediction Model and Their Standard Errors

Risk factors	**Estimates** $\hat{\beta}$	**Estimated Standard Errors**
Intercept	-6.5181	0.5800
SEX (F)	-0.3588	0.0753
DAILY SMK (NO)	-0.2841	0.0708
HIGH SCH + (NO)	0.2205	0.0724
DIABETES (NO)	-0.7335	0.0925
HIGH BP (NO)	-0.6987	0.0757
BMI	0.0154	0.0077
AGE	0.1017	0.0154

We illustrate prediction of a probability of having CHD for a 54 year-old man who is not a current or former daily smoker, who has education higher than high school, who is diabetic, but does not have the high blood pressure, and whose BMI is equal to 22.5. First, the predicted risk score $\hat{\eta}$ is equal to -2.572 with its variance estimated as 0.289. The predicted probability of having CHD for this person is 0.071 with variance estimated by the bootstrap estimator (4) as 0.0007. The confidence interval for this person's probability based on transformed individual risk score limits is (0.0259, 0.1797).

The predicted population average probability of having CHD in Ontario is $\hat{p}\left(\hat{\beta}\right)$=0.047.

We apply the prediction model estimated using the sample (training sample) from Ontario to predict the population average risk of having CHD in Québec using the independent Québec portion (*n*=27,309) of the CCHS Cycle 3.1 sample (validation sample). The population average probability is predicted as 0.050.

The components of the estimated variances of the predicted population average risks for both provinces are given in Table 2.

Table 2. Prediction of Population Average Risk for Ontario and Québec Based On The Ontario (Training) Sample:

	Ontario (n=37,310)	**Québec (n=27,309)**
Average risk $\hat{p}(\hat{\beta})$	**0.047**	**0.050**
Variance estimates (6) and (16)	0.022 x 10^{-4}	0.029 x 10^{-4}
Components:		
Estimate of E_1V_2 (11)		0.002 x 10^{-4}
Estimate of V_1E_2 (14)		0.027 x 10^{-4}
Alternate variance estimate (17)		0.028 x 10^{-4}
Components:		
Estimates of E_1V_2 (7) and (10)	0.001 x 10^{-4}	0.002 x 10^{-4}
Estimate of V_1E_2 (15)		0.026 x 10^{-4}
95% confidence interval	**(0.0441, 0.0500)**	**(0.0467, 0.0533)**

If we assume that β is known, the larger component of the variance, $\hat{V}_1(\hat{E}_2)$, is set equal to zero and the variance of the average probability is seriously underestimated.

6. Conclusion

When predicting the risk probabilities of having or developing a disease using a model estimated from the sample, the sampling variability has to be accounted for. In this paper we presented a variance estimation approach using the survey bootstrap method (Rao, Wu, Yue, 1992) which is the method of choice for many of Statistics Canada's analytic surveys. The method itself relies on the availability of a large number of sets of properly obtained bootstrap survey weights. These weights are available for many of Statistics Canada's analytic surveys such as the CCHS whose data were used for the illustration.

The illustration (Table 2) shows also that the sampling variability in the estimates of the beta coefficients is the major component of the variance of the average probability prediction and should not be neglected, as is sometimes done when the estimated coefficients in the prediction model are treated as constants.

Note that there are other ways to compute the variances of the statistics explored in this paper, most notably the Taylor linearization method. However, currently, only the resampling methods incorporate the most of survey information (non-response adjustments, post-stratification, etc.) into the variance estimation. On the other hand, the properties of the bootstrap variance estimators in the context of prediction of probabilities need to be more thoroughly investigated theoretically and by means of empirical studies

References

Binder, D., Kovacevic, M., and Roberts, G. (2005). Design-Based Methods For Survey Data: Alternative Uses Of Estimating Functions. Proceedings of the Survey Research Methods Section, ASA.

D'Agostino R.B., Vasan R.S., Pencina M.J., Wolf, P.A., Cobain, M., Massaro, J.M., Kannel, W.B. (2008) General cardiovascular risk profile for use in primary care. The Framingham Heart Study. *Circulation*. Vol.117(6): 743-753.

Rao, J.N.K. and Wu, C.F.J.(1988). Resampling inference with complex survey data. *J. American Statistical Association*, 83, 231-241

Rao,J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

Rosella,L., Manuel, D., Stukel, T. (2008). Predicting New Cases Of Diabetes In Canada: Historical And Future Trends Using A Population Based Risk Tool. Presented at the *Joint Annual Conference of the Statistical Society of Canada and the Sociéte Française de Statistique,* Ottawa, May 25-29, 2008.

Rosella,L. Manuel, D., Burchill,C., Stukel, T. (2008). Predicting New Cases Of Diabetes In Canada. *Manuscript. Submitted for publishing*

Rust, K. and Rao, J.N.K. (1996). Variance Estimation for Complex Surveys Using Replication Techniques. *Statistical Methods in Medical Research*, Vol. 5, No. 3, 283-311.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.