Published in final edited form as:

Stat Med. 2015 July 10; 34(15): 2353–2367. doi:10.1002/sim.6498.

The Impact of Covariate Measurement Error on Risk Prediction

Polyna Khudyakov[a,*], **Malka Gorfine**[b], **David Zucker**[c], and **Donna Spiegelman**[d]

[a]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A.

[b]Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel

[c]Department of Statistics, Hebrew University of Jerusalem, Mt. Scopus, Jerusalem, Israel

[d]Departments of Epidemiology, Biostatistics, Nutrition and Global Health, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A.

Abstract

In the development of risk prediction models, predictors are often measured with error. In this paper, we investigate the impact of covariate measurement error on risk prediction. We compare the prediction performance using a costly variable measured without error, along with error-free covariates, to that of a model based on an inexpensive surrogate along with the error-free covariates. We consider continuous error-prone covariates with homoscedastic and heteroscedastic errors, and also a discrete misclassified covariate. Prediction performance is evaluated by the area under the receiver operating characteristic curve (AUC), the Brier score (BS), and the ratio of the observed to the expected number of events (calibration). In an extensive numerical study, we show that (i) the prediction model with the error-prone covariate is very well calibrated, even when it is mis-specified; (ii) using the error-prone covariate instead of the true covariate can reduce the AUC and increase the BS dramatically; (iii) adding an auxiliary variable, which is correlated with the error-prone covariate but conditionally independent of the outcome given all covariates in the true model, can improve the AUC and BS substantially. We conclude that reducing measurement error in covariates will improve the ensuing risk prediction, unless the association between the error-free and error-prone covariates is very high. Finally, we demonstrate how a validation study can be used to assess the effect of mismeasured covariates on risk prediction. These concepts are illustrated in a breast cancer risk prediction model developed in the Nurses' Health Study.

Keywords

risk prediction; probit regression; logistic regression; measurement error; ROC-AUC; Brier score

1. Introduction

Risk prediction models are used to translate research findings into valuable tools to assist prognostic assessment, screening algorithms, and clinical decision making. Several widely

[*] Correspondence to: Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A. polyna.khudyakov@channing.harvard.edu.

used prognostic tools have been developed from epidemiologic data [1]–[3]. In epidemiology, key risk factors are often measured with error [4]-[15] and a great variety of statistical and epidemiological literature has dealt with correcting bias in *relative* risk estimators caused by measurement error [16]–[18]. However, little attention has been paid to the impact of measurement error on prediction, and a small number of papers, that considered influence of measurement error on prediction, analyzed it in terms of effect on risk estimates [19]-[21]. Some discussion of methods and problems in prediction in the presence of measurement error can be found in [17]–[18]. What appears to be the prevailing view was expressed by Carroll et al. ([17], Sec. 2.6): “Generally, there is no need for the modeling of measurement error to play a role in the prediction problem. If a predictor X is measured with error and one wants to predict a response based on the error-prone version W of X, then except for a special case, it rarely makes any sense to worry about measurement error. The reason is quite simple: W is error-free as a measurement of itself!” This means that a regression model fitted with noisy covariates can be a valid prediction model. A valid, or perfectly calibrated, model is unbiased in the sense that, on average, there is perfect agreement between the observed outcomes and predictions. For example, if the model predicts that 10% of those at risk will develop a certain disease, the observed frequency of the disease will be, on average, 10%. Formally, a model is perfectly calibrated, or unbiased, if $E(\hat{p_i}) = E(Y_i|X_i, Z_i)$, where $\hat{p_i}$ is the estimated conditional probability of $\{Y_i = 1\}$ given the covariates. An exception noted by Carroll et al. [17] occurs when the prediction model is estimated from a study with a given amount or form of covariate error but is to be applied to a population where the amount or form of covariate error is different. That is, when measurement error is not “transportable”, a risk prediction model perfectly calibrated to a population with one underlying error structure, will not be well calibrated to a population with another underlying error structure. This problem was recently addressed for replicated covariates measured under an independent additive error model [22].

Although mismeasured covariates may provide valid predictions, sometimes their respective costly error-free measures are, or could be, available. The main goal of this paper is to investigate the potential gain in prediction performance from using the error-free covariates. For example, in the Nurses’ Health Study (NHS) [23], a risk prediction model for 20-year breast cancer incidence was considered given a set of well known risk factors, some of which are error-free while others are error-prone. Risk factors in this model include age, age at menarche, age at first birth, family history of breast cancer, number of past breast biopsies, as in [2] and in [24]-[26], and could potentially be enhanced by data on alcohol intake [27, 28] and $\alpha-$carotene intake [14, 29]. Here, the main study consists of 68,555 female nurses whose alcohol and $\alpha-$carotene intake were assessed through a self-administered 131-item food frequency questionnaire (FFQs) at baseline in 1986. The FFQ measures dietary intake with moderate to substantial error [4]-[14], [30]. The validation study was conducted in NHS in 1986 [14] among 191 NHS cohort members by the corresponding values obtained two 7-day of weighed diet records for dietary intake assessed by FFQ. Using the methodology developed in this paper, we compared two risk prediction models, one using the surrogate intake measures obtained from the FFQs and the other using the gold standards from diet records, to assess the potential gain in risk prediction performance due to the availability of perfectly measured predictors.

Among the generalized linear regression models for binary outcomes, logistic and probit are the two most widely considered link functions. Cox and Snell ([31], pp. 23) argued that although the two models produce different parameter estimates, these estimates usually end up with similar standardized impacts of the covariates; Greene ([32], p. 875) concluded that "in most applications, it seems not to make much difference"; Gill ([33], p. 33) indicated that they "provide identical substantive conclusions"; and similar conclusions appears regularly in the discussions comparing between the logit and probit models (e.g. [34]- [37]). Since the probit model provides analytically tractable expressions, in contrast to the logit model, we focus on probit models.

This paper investigates the effect of measurement error on risk prediction and demonstrates how a validation study can be used to assess the effect of mismeasured covariates on risk prediction. Specifically, we consider a probit regression model to predict a binary outcome based on error-free and error-prone covariates; linear measurement error models with homoscedastic and heteroscadastic error; and varied degrees of measurement error between the error-free and error-prone covariates. The case of a misclassified binary covariate is studied as well. The settings with heteroscadastic error variance or misclassified covariates are no longer follow the probit model. Hence, in these scenarios we use also the generalized additive model (GAM) of Hastie and Tibshirani [38].

The performance of the risk prediction models is evaluated in terms of the following three criteria [39]: discrimination,prediction accuracy, and calibration. Discrimination is assessed using the area under the receiver operating characteristic curve (AUC). Prediction accuracy (overall performance) is assessed using the Brier score (BS). Calibration is assessed by examining the ratio between the overall observed number of events and the expected number of events under the model. Although the AUC and BS are closely related quantities and under certain assumptions have an explicit relationship [40]-[41], since both are commonly used to evaluate model prediction quality and because the magnitude of comparisons differ between these two measures, we report results for both of them. For clarity of presentation of the effect of measurement error on risk prediction, we assume that the true values of the probit models parameters are available.

This article is organized in five sections. Section 2 defines the probit outcome model and the two measurement error models we consider in this paper, and several common measures to be used for comparing the performance of various prediction models to be examined. Section 3 provides a detailed description of the numerical study conducted for investigating the impact of measurement error on the performance of the prediction models. This investigation includes homoscedastic and heteroscedastic error models for a continuous covariate, and also models with a misclassified binary covariate . This section concludes with a summary of the results from the numerical study. In Section 4, we demonstrate the use of a validation study to assess the effect of mismeasured covariates on risk prediction, through a breast cancer risk prediction model in the Nurses' Health Study [23]. Finally, Section 5, summarizes the practical implications of this work.

2. Models and methods

2.1. The models

Consider a regression model relating a binary response variable Y to a true predictor X and an error-free predictor Z. Let W denote the error-prone value corresponding to X. We define the binary outcome Y in terms of a continuous latent random variable Y_0. Specifically, let $Y = I(Y_0 \geq 0)$, with

$$Y_0 = \alpha + \beta X + \gamma Z + \eta \quad (1)$$

where α, β and γ are unknown regression coefficients and η is a normal random variable with zero mean and unknown variance ω^2. In addition, we assume that (Y_0, X, Z) is multivariate normally distributed with expectation $(\alpha, 0, 0)^T$, and the following variance-covariance components:

$$\begin{aligned} var\left(Y_0\right) &= \beta^2\sigma_X^2 + \gamma^2\sigma_Z^2 + 2\beta\gamma\rho_{X,Z}\sigma_X\sigma_Z \\ &+ \omega^2; var\left(X\right) \\ &= \sigma_X^2; var\left(Z\right) \\ &= \sigma_Z^2; cov\left(Y_0, X\right) \\ &= \beta\sigma_X^2 + \gamma\rho_{X,Z}\sigma_X\sigma_Z; \rho_{X,Z} \\ &= cov\left(X, Z\right) / \left(\sigma_X\sigma_Z\right) \end{aligned}$$

; and $cov\left(Y_0, Z\right) = \gamma\sigma_Z^2 + \beta\rho_{X,Z}\sigma_X\sigma_Z$. This model is known in quantitative genetics as a liability model (see for example [42]). The liability model assumes that a normal random variable Y_0 is related to the discontinuous trait Y by a threshold, providing a convenient methodological framework, as here. However, we note that the results provided in this paper are based on the model (1). For more general conclusions, further numerical studies are needed.

Next, we assume that W is a surrogate for X, i.e. the conditional distribution of Y given (X, Z, W) equals the conditional distribution of Y given (X, Z). In addition, we assume the linear regression measurement error model $W = c + dX + \varepsilon$, where ε is independent of X and is normally distributed with mean zero. This measurement error model is an established model in the measurement error modeling literature [17, 43, 44]. The special case $c = 0$ and $d = 1$ yields the classical additive measurement error model $W = X + \varepsilon$. Under the homoscedastic measurement error model, it is assumed that ε has constant variance σ_ε^2. The heterosocedastic setting will specify a conditional normal distribution of ε given the true covariate X such that the conditional variance of ε is a function of X.

Let Φ denote the cumulative distribution function of the standard normal distribution. The true and surrogate probit prediction models are based on (Y, X, Z) and (Y, W, Z), respectively. Then,

$$Pr\left(Y=1|X, Z\right) = \Phi\left\{\left(\alpha + \beta X + \gamma Z\right)/\omega\right\}, \quad (2)$$

and under the linear regression measurement error model and the multivariate distribution of (Y_0, X, Z) described above

$$Pr\left(Y=1|W,Z\right)=\Phi\left\{\left(\tilde{\alpha}+\tilde{\beta}W+\tilde{\gamma}Z\right)/\tilde{\omega}\right\}. \quad (3)$$

Based on the attenuation matrix ([17], p. 53, Eq. (3.12), [43]-[45]), $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$ and $\tilde{\omega}$ can be explicitly written as a function of α, β, γ, $\rho_{X,Z}$, *c*, *d*, σ_ε^2, σ_Z^2, and σ_X^2. The exact form of $\left(\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\omega}\right)$ for the scenarios considered in this paper will be given in Section 3.

2.2. Performance measures

Let φ denote the density of the standard normal distribution, $f_S(\cdot)$ the density of $S = \alpha + \beta X + \gamma Z$, and $f_{S0}(\cdot)$ and $f_{S1}(\cdot)$ the conditional densities of S given the events $\{Y = 0\}$ and $\{Y = 1\}$, respectively. Then, by Bayes rule we obtain $f_{S0}(s) = f_S(s)\ \Pr(Y_0 \le 0|S = s)/\ \Pr(Y_0 \le 0)$ and $f_{S1}(s) = f_S(s)\ \Pr(Y_0 > 0|S = s)/\ \Pr(Y_0 > 0)$. Hence, $f_{S_0}\left(s\right)=\tau^{-1}\phi\left\{\left(s-\alpha\right)/\tau\right\}\left\{1-\Phi\left(s/\omega\right)\right\}/\left\{1-\Phi\left(\alpha/\sqrt{\omega^2+\tau^2}\right)\right\}$ and $f_{S_1}\left(s\right)=\tau^{-1}\phi\left\{\left(s-\alpha\right)/\tau\right\}\Phi\left(s/\omega\right)/\left\{\Phi\left(\alpha/\sqrt{\omega^2+\tau^2}\right)\right\}$, where $\tau^2=\beta^2\sigma_X^2+\gamma^2\sigma_Z^2+2\beta\gamma\rho_{X,Z}\sigma_X\sigma_Z$. Since, $AUC(X, Z) = P\ (S_0 < S_1)$ [48], it can be shown that under the probit model (2),

$$AUC\left(X,Z\right)=\int_{-\infty}^{\infty}\int_{s_0}^{\infty}f_{S_0}\left(s_0\right)f_{S_1}\left(s_1\right)ds_0ds_1. \quad (4)$$

The AUC varies between 0.5 and 1.0 with higher value indicating a better discriminative model.

The Brier Score (BS) [46] quantifies the overall model performance and consists of the squared distance between the actual binary outcome, Y , and the estimated conditional probability of $\{Y = 1\}$ given the covariates, denoted by $\hat{p}$. The Brier Score is defined as $BS(X, Z) = E\{(Y - \hat{p})^2\}$. Under the probit model (2), it is easy to verify that

$$BS\left(X,Z\right)=\Phi\left(\alpha/\sqrt{\omega^2+r^2}\right)-\int_{-\infty}^{\infty}\tau^{-1}\phi\left\{\left(s-\alpha\right)/\tau\right\}\Phi^2\left(s/\omega\right)ds. \quad (5)$$

The smaller the BS is, the better the prediction.

The AUC and BS of the surrogate model (3), under homoscedastic measurement error $AUC(W, Z)$ and $BS(W, Z)$, can be derived similarly. All of the above integrals can be calculated using numerical quadrature routines, which are available in variety of software packages. However, these calculations are time consuming.

Alternatively, the AUC and BS can be calculated through Monte Carlo simulation. Specifically, a large random sample of $(Y_{01}, X_1, Z_1), \ldots, (Y_{0n}, X_n, Z_n)$ is generated from the multivariate normal distribution described above. Let $Y_i = I(Y_{0i} \ge 0)$, $i = 1, \ldots, n$; $\hat{p}_i$ be the estimated conditional probability of $\{Y_i = 1\}$ given the covariates; $n_1=\sum_{i=1}^{n}I\left(Y_i=1\right)$; and $n_0 = n - n_1$. A nonparametric estimator of the AUC ([47], p. 493) for predicting the event $\{Y = 1\}$ is defined by

$$(n_0 n_1)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} U\left(\hat{p}_i, \hat{p}_j\right) Y_i \left(1 - Y_j\right), \quad (6)$$

where *U*(*a*, *b*) = *I*(*a* > *b*) + 0.5*I*(*a* = *b*). A natural empirical estimator of BS [39] is given by

$$n^{-1} \sum_{i=1}^{n} (\hat{p}_i - Y_i)^2. \quad (7)$$

In a series of settings, we contrasted the theoretical AUC and BS (4)-(5) calculated by numerical quadrature to the Monte Carlo simulation approach using the above estimators (6)-(7) and a very large sample size. As expected, we found that these methods produced almost identical results. Hence, to save computing time, from now on we will use the simulation-based approach.

Calibration performance was evaluated by the ratio of the observed and the expected number of events $\mathrm{O/E} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} \hat{p}_i$ [39]. If the model is well calibrated, we expect O/E to be close to 1. Since Carroll et al. [17] argued that a surrogate model is valid but omitted a formal proof, we present one by showing that the expected number of events under the true and the surrogate models are equal:

$$\begin{aligned} E\left\{Pr\left(Y=1|W,Z\right)\right\} = \; & E\left[E\left\{I\left(Y_0 \geq 0\right)|W,Z\right\}\right] \\ = \; & E\left(E\left[E\left\{I\left(Y_0 \geq 0\right)|X,W,Z\right\}|W,Z\right]\right) \\ = \; & E\left(E\left[E\left\{I\left(Y_0 \geq 0\right)|X,Z\right\}|W,Z\right]\right) \\ = \; & E\left[E\left\{Pr\left(Y=1|X,Z\right)|W,Z\right\}\right] \\ = \; & E\left\{Pr\left(Y=1|X,Z\right)\right\}. \end{aligned} \quad (8)$$

The third equality holds due to the surrogacy assumption, and the explanation of the last transition is presented below

$$\begin{aligned} E\left[E\left\{Pr\left(Y=1|X,Z\right)|W,Z\right\}\right] = \; & \int\int\int \Phi\left(\left(\alpha+\beta x+\gamma z\right)/\omega\right) f_{X|W,Z}\left(x|w,z\right) f_{W,Z}\left(w,z\right) dwdxdz \\ = \; & \int\int \Phi\left(\left(\alpha+\beta x+\gamma z\right)/\omega\right) \left(\int f_{X,W,Z}\left(x,w,z\right) dw\right) dxdz \\ = \; & \int\int \Phi\left(\left(\alpha+\beta x+\gamma z\right)/\omega\right) f_{X,Z}\left(x,z\right) dxdz \\ = \; & E\left\{Pr\left(Y=1|X,Z\right)\right\}. \end{aligned}$$

If the true model of Pr(*Y* = 1*[notdef]W*, *Z*) is used, numerical studies of calibration in the presence of measurement error in *X* are not needed since perfect calibration is established theoretically above. However, in the presence of binary covariate misclassification and in the presence of heteroscedastic error in continuous *X*, the linear model in the probit link will be mis-specified, thereby might leading to imperfect calibration of some degree to be determined below.

3. Numerical study

The main goal of this section was to study the difference in prediction performance of the true model based on (*Y*, *X*, *Z*) compared to the surrogate model based on (*Y*, *W*, *Z*). Results from the homoscedastic error setting are given in Section 3.1. Several heteroscedastic error settings and a binary covariate subject to misclassification are described in Sections 3.2 and 3.3, respectively.

3.1. Homoscedastic error model

The relationships among Y_0, X and Z will have a critical impact on the differences in performance between the true and surrogate models. The following parametrization allows one to systematically explore a wide range of associations among the variables. Let

$$\begin{aligned} \alpha= & \ \Phi^{-1}(p), \\ \beta= & \ \sqrt{R^2}\, sin\, \varphi / \sqrt{1+\rho_{X,Z}\, sin\ (2\varphi)}, \\ \gamma= & \ \sqrt{R^2}\, cos\, \varphi / \sqrt{1+\rho_{X,Z}\, sin\ (2\varphi)}, \end{aligned}$$

where $\varphi \in [0, \pi]$, $p \in (0, 1)$, $R^2 = 1 - \omega^2$, $\omega^2 = var(\ )$, so that R^2 equals the proportion of the total variation in Y_0 explained by the linear regression model with (X, Z). Here, the expectations of X, Z and W were set to 0 and the variances of X and Z were set to 1. Then, Y_0 is normally distributed with mean α and variance 1; φ describes the relative importance of X and Z as predictors of Y_0; and the parameter p satisfies $p = \Pr(Y = 1)$. It can be shown that the parameters of the surrogate model (3) are

$$\tilde{\beta}=\beta d\left(1-\rho_{XZ}^2\right)/\left(d\left(1-\rho_{XZ}^2\right)+\sigma_\varepsilon^2\right), \quad (9)$$

$$\tilde{\gamma}=\left[\gamma d^2\left(1-\rho_{XZ}^2\right)+\beta\rho_{XZ}\sigma_\varepsilon^2+\gamma\sigma_\varepsilon^2\right]/\left(d\left(1-\rho_{XZ}^2\right)+\sigma_\varepsilon^2\right), \quad (10)$$

$$\tilde{\alpha}=\alpha-\tilde{\beta}c, \quad (11)$$

and

$$\tilde{\omega}^2=1-\left(d^2+\sigma_\varepsilon^2\right)\tilde{\beta}^2-\tilde{\gamma}^2-2\tilde{\beta}\tilde{\gamma}d\rho_{XZ}. \quad (12)$$

Since $\sin^2 \varphi + \cos^2 \varphi = 1$ and $\sin(2\varphi) = 2 \sin \varphi \cos \varphi$, $\tau^2 = R^2$ and $AUC(X, Z)$ and $B(X, Z)$ are constant functions of $\rho_{X,Z}$. Obviously this is not the case for $AUC(W, Z)$ and $BS(W, Z)$, as will be demonstrated in Section 3.4.

In this numerical study, we focused on the classical measurement error model $W = X + \varepsilon$, i.e., the special case $c = 0$ and $d = 1$. We considered the following values for the parameters: $R^2 = 0.1, 0.3, 0.5, 0.7, 0.9$, $p = 0.05, 0.15, 0.3, 0.5$, $\varphi = 0, \pi/36, 2\pi/36, \ldots, \pi$, $\rho_{X,Z} = -0.9, -0.5, -0.3, 0, 0.3, 0.5, 0.9$ and $\sigma_\varepsilon^2=0.5, 1, 2$. The values of σ_ε^2 are motivated by the NHS validation data where σ_ε^2 is about 1.7 times the variance of the error-free X in the case of α-carotene and about 0.4 times the variance of the error-free X in the case of alcohol intake (see Table 5). The simulation results are based on one sample of size of 1,000,000 observations for each configuration. The sample size is big because the purpose of this simulation study is calculating the values of the performance measures under different models, and not to study their finite sample properties.

In addition, we investigated the possible improvement in prediction performance in a surrogate model in which an auxiliary variable (AV) is added to the mean for the regressor X. An AV (also known as an instrumental variable) is one that is associated with X but not with Y_0 given X and Z [17]. Hence, we defined the AV, denoted by V , as $V = \rho_{X,V} X + \xi$ where ξ is a zero-mean normally distributed random variable with variance $1 - \rho^2_{X,V}$, and ξ is independent of all other random variables in the model. We considered $\rho_{X,V} = -0.9, -0.5, -0.3, 0.3, 0.5, 0.9$ and compared the prediction performance of the true model with regressors (X, Z) to models with regressors (W, Z) and (V, W, Z).

For each scenario, we estimated the AUC, BS and O/E in five models: (i) with X as a single covariate; (ii) with Z as a single covariate; (iii) with (X, Z) as covariates; (iv) with (W, Z) as covariates; and (v) with (V, W, Z). The results are presented in Section 3.4.

3.2. Heteroscedastic error model

The heteroscedastic error models were modifications of the models in Section 3.1, such that ε given X is a zero-mean normally distributed random variable with variance that depends on X. Three models for heteroscedastic error were studied: (I) $var(\varepsilon|X) = 1.3|X|$; (II) $var(\varepsilon|X) = \exp(0.2X)$; (III) $var(\varepsilon|X) = 1.7|\sin(2X)|$. Similar scenarios were considered in [20]. In each of these models, the marginal error variance equaled 1, i.e. $\sigma^2_\varepsilon = 1$, and $\rho_{X,Z} = 0$. Figure 1 illustrates the association of X and the error variate ε under the above heteroscedastic error models.

In contrast to the homogeneous error variance setting, $P(Y = 1|W, Z)$ no longer follows the probit model. Hence, we used also the generalized additive model (GAM) of Hastie and Tibshirani [38], as a flexible approximation to the true model. Our use of the GAM approach is motivated by the idea that if heteroscedasticity were present, the analyst building the prediction model would notice departures from linearity on the probit scale and would therefore fit a more flexible model. The generalized additive model incorporates this flexibility into the setting of generalized linear models. We carried out the GAM fitting using the "gam" function in the R package "mgcv" [49]-[50]. The function "gam" was applied with the default parameters, using penalized regression splines with smoothing parameters selected by the generalized cross validation criterion [51].

3.3. Binary covariates with misclassification

For the case of a dichotomous predictor, we denote the true binary error-prone exposure by $\tilde{X}$, its binary surrogate as $\tilde{W}$, and an error-free binary covariate $\tilde{Z}$, such that $\tilde{X} = I(X > 0)$ and $\tilde{Z} = I(Z > 0)$, with X and Z independently normally distributed with means zero and variances 1. The latent variable Y_0 was defined as $Y_0 = \alpha + \beta\tilde{X} + \gamma\tilde{Z} = \eta$, where η is zero-mean normally distributed random variable with variance 1. We considered the case with $\alpha = -1$, and $\beta = \gamma = 1$. Denote the misclassification probabilities as $\Pr(\tilde{W} = 1|\tilde{X} = 0) = q_{10}$ and $\Pr(\tilde{W} = 0|\tilde{X} = 1) = q_{01}$. The results presented in the section that follows are with $q_{10} = q_{01} = q$ and $q = 0.05, 0.15, 0.25$ or 0.5. In this case, the probit model for $\Pr(Y = 1|W, Z)$ is misspecified, since in fact $\Phi^{-1}\{\Pr(Y = 1|W, Z)\}$ is not linear in Z. The misclassification rates are designed to cover the entire range, and the high misclassification rates are motivated by previous analysis of FFQ data [29], where q was close to 0.5. As previously, the results are based on probit models.

3.4. Main results

Table 1 presents the $AUC(X, Z)$ and $BS(X, Z)$ for various values of R^2 and $p = \Pr(Y = 1)$, as it is well-known that prediction performance measures depend on the prevalence of the event of interest in the population [52]. As expected, the AUC value decreases and the BS value increases as R^2 decreases and also as p increases. The last finding is due to fact that α is an increasing function of p, $\alpha = \Phi^{-1}(p)$, and as α increases the harder it is for the covariates (X, Z) to discriminate between events and non-events.

To exemplify the general pattern of the prediction performances as a function of the regression coefficients, we present the cases with $R^2 = 0.5$ and $p = 0.5$ in Figures 2-4, and the AUC and BS results for cases with $R^2 = 0.1, 0.3$ and $p = 0.1$ are presented in the Supporting information, Figures S1-S4. The results for the AUC and BS in the homoscedastic setting are given in Figures 2 and 3, respectively. Each figure provides the results for various values of $\rho_{X,Z}$. The horizontal axis of each plot represents ϕ, which defines the relative importance of X and Z in model (1). For ease of interpretation we added in the horizontal axes the respective values of β and γ. Each plot consists of $AUC(X, Z)$, $AUC(Z)$, and $AUC(W, Z)$ with $\sigma_\varepsilon^2=0.5, 1, 2$. As expected, $AUC(Z) \le AUC(W, Z) \le AUC(X, Z)$ and $AUC(W, Z)$ decreased as σ_ε^2 increased for fixed β and γ. Under $\rho_{X,Z} = 0$, the minimal values of $AUC(W, Z)$ and $AUC(Z)$, for a given σ_ε^2 are attained at γ = 0, namely, when Z makes no contribution to the prediction model. When $\rho_{X,Z} \neq= 0$, the minimal values of $AUC(W, Z)$ and $AUC(Z)$ are attained at different values of ϕ due to the effect of the correlation between X and Z. With a strong association such as $\rho_{X,Z} = -0.9$, the maximum differences between $AUC(X, Z)$ and $AUC(W, Z)$ could be large and are attained only when β and γ are approximately equal. For example, if $\sigma_\varepsilon^2=2$, $\phi = 45°$, $AUC(X, Z) - AUC(W, Z) = 0.256$. As the differences between β and γ increased, the differences between the various AUCs decreased. In particular, when $\beta = \gamma - \approx 0.36$, the differences between the AUCs are almost zero. As the association of X and Z decreased, the differences between the AUCs were substantial under a wide range of ϕ, namely, under wide ranges of β and γ. In particular, under $\rho_{X,Z} = 0.3$, $\sigma_\varepsilon^2=2$, $AUC(X, Z) - AUC(W, Z)$ is at least 0.15 for all $\phi \in [70°, 120°]$, where 0.15 is a substantial decrease in the AUC due to error.

For the BS results (Figure 3), similar patterns as found for the AUC were observed, $BS(X, Z) \le BS(W, Z) \le BS(Z)$. When X and Z were at most moderately associated (i.e. $\rho_{X,Z} \le 0.5$), the accuracy loss due to the mismeasured W was substantial under almost the entire range of ϕ. For example, with $\rho_{X,Z} = 0.3$ and $\sigma_\varepsilon^2=2$, $BS(W, Z) - BS(X, Z)$ was as high as 0.07.

In summary, when the measurement error was of small magnitude, the discrimination and accuracy of the prediction was decreased minimally. However, with moderate or severe measurement error, the decline in AUC and BS from using (W, Z) instead of (X, Z) was dramatic. It should be again noted that because the probit model in (W, Z) was perfectly specified, these observed declines in the quality of the prediction were due exclusively to increased variability in the prediction and not to an increase in bias.

The contribution of the AV, V , to the prediction model performance, under the homoscedastic measurement error model with $R^2 = 0.5$ and $p = 0.5$ is given in Figure 4, in terms of AUC. The results of BS are presented in the Supporting information Figure S5. Since the AUC and BS are symmetric for $\rho_{X,Z}$ and $-\rho_{X,Z}$, and since the results are the same for $\rho_{X,V}$ and $-\rho_{X,V}$, we present only the case where $\rho_{X,Z} > 0$ and $\rho_{X,V} > 0$. The left and right columns of Figure 4 are with correlation coefficient of X and V of 0.9 and 0.5, respectively ($\rho_{X,V} = 0.3$ is omitted). As expected, for any σ_ε^2, $AUC(W, Z, V) \geq AUC(W, Z)$. Contrasting $AUC(W, Z, V)$ with $AUC(W, Z)$, revealed that adding an AV to the model could considerably improve the AUC and BS under moderate or substantial measurement error, as long as there was at most a moderate association of X with Z. When $|\rho_{X,Z}|$ was high, the added value of including V in the model was very small. For example, for $\rho_{X,V} = 0.9$, $\rho_{X,Z} = 0.9$, $\sigma_\varepsilon^2=2$ and $\varphi = 45°$, we get $AUC(W, Z) = 0.825$, $AUC(W, Z, V) = 0.829$, $BS(W, Z) = 0.171$ and $BS(W, Z, V) = 0.169$; and for a similar scenario but with $\rho_{X,Z} = 0.3$ and $\varphi = 80°$, we get $AUC(W, Z) = 0.662$, $AUC(W, Z, V) = 0.801$, $BS(W, Z) = 0.198$ and $BS(W, Z, V) = 0.173$.

The results from the heteroscedastic settings with $R^2 = 0.5$, $p = 0.5$, and $\rho_{X,Z} = 0$ are presented in Tables 2-3. It is apparent that the AUC and BS values from Scenarios (I)-(III) were very close to their respective AUC and BS values from the corresponding homoscedastic error models with $\rho_{X,Z} = 0$ and $\sigma_\varepsilon^2=1$. The conclusions from the homoscedastic and heteroscedastic settings were the same, in the sense that, for example, $AUC(Z) \leq AUC(W, Z) \leq AUC(X, Z)$ and $AUC(W, Z)$ decreased as σ_ε^2 increased for fixed β and γ. Moreover, in Tables 2-3 we contrasted the GAM and probit models in terms of AUC and BS. The similarity between these results indicates that in all the settings considered, misspecifying the form of the prediction model had little impact.

The results of probit model with binary regressors in the misclassification setting are summarized in Table 4. As with continuous predictors, $AUC(\tilde{Z}) \leq AUC(\tilde{W}, \tilde{Z}) \leq AUC(\tilde{X}, \tilde{Z})$. As expected, $AUC(\tilde{W}, \tilde{Z})$ decreased and $BS(\tilde{Z}, \tilde{W})$ increased as q increased. Also, for $q = 0.5$, $AUC(\tilde{W}, \tilde{Z}) = AUC(\tilde{Z})$, because here $\tilde{W}$ contains no information about $\tilde{X}$. As before, the gain in prediction performance due to the use of $\tilde{X}$ instead of the mismeasured $\tilde{W}$ was substantial, especially when X and Z were moderately or weakly associated. For example, with $\rho_{X,Z} = 0.7$ and $q = 0.25$, $AUC(\tilde{X}, \tilde{Z}) - AUC(\tilde{W}, \tilde{Z}) = 0.019$ and with $\rho_{X,Z} = 0.3$ and $q = 0.25$, $AUC(\tilde{X}, \tilde{Z}) - AUC(\tilde{W}, \tilde{Z}) = 0.047$. An improvement of AUC of approximately 0.05 is more than a modest improvement [53]–[54].

Under the homoscedastic, heteroscedastic and misclassification error models, the ratio of the observed and expected number of events were always close to one (between 0.9994 and 1.0003), for all prediction models considered. Under the homoscedastic settings, the results are as expected, given (8). The results from the other settings, when the probit model is mis-specified, imply that, asymptotically, the surrogate model is approximately valid even when the prediction model is mis-specified and is being approximated by a probit model or GAM.

4. An illustrative example

In this example, we illustrate how to calculate the extent to which the prediction would be improved if *X* were available rather than the surrogate measurement, *W* . We evaluated prediction models in the Nurses' Health study (NHS) [23] for breast cancer incidence using the error-prone covariates for alcohol (g/day) [27, 28] and α-caratenoid consumption (mg/day) [29] and an error-free covariate, the Gail score [2]. With data from the NHS dietary validation studies [14] and the theory presented in this paper, we compared the AUC and BS from the model with surrogate dietary measurements to the possible values that we could have obtained if the measurements had been available in the main study.

The NHS was established in 1976, when 121,700 female U.S. registered nurses between the ages of 30 and 55 years responded to a mailed study questionnaire on medical history and lifestyle. Subsequent questionnaires have been mailed every 2 years. Further details on the study, including information on disease confirmation, have been published elsewhere [23]. We restricted our analyses to the 20-year period from 1986 to 2006, and excluded women with a history of cardiovascular disease or cancer, including lobular or ductal breast carcinoma *in situ* at baseline. We were then left with a cohort of 66,346 women aged 40-71 years in 1986, within which 3,065 women developed invasive breast cancer within 20 years of the return date of their 1986 questionnaire. The validation study assessed consumption from a self-administrated food frequency questionnaire and two 7-day diet records completed approximately 6 months apart (for more details see [14]).

We considered the effects of the error-prone predictors in various scenarios: (i) α-carotene alone, (ii) alcohol consumption alone and (iii) α-carotene and alcohol consumption, and analyzed models containing the error-prone covariates alone and together with the error-free covariate, the Gail score. In total, we studied six models. The Gail score was calculated from the algorithm and parameters of Gail et al.'s model [2]. The means and variances of the regressors based on the validation study are given in the top part of Table 5. The Pearson correlation between the Gail score and α-carotene was 0.124; between the Gail score and alcohol was -0.036; and between α-carotene and alcohol was 0.030. The estimated regression coefficients of the models using the main study are given at the bottom of Table 5.

The main results of this analysis are given in Table 6. We randomly split the main study into two datasets of equal sizes. One set was used for estimating the regression parameters, and the other for estimating the AUC and BS based on (6)-(7). We compared the surrogate performance measures of probit, logit and GAM models. It is evident that the AUCs and BSs of the probit and logit models are very similar, and as expected, the AUC and BS are insensitive to the choice of the logit or probit link function. Also the -2log-likelihood (-2LL) measures of the probit and logit models are very similar. Moreover, the AUCs and BSs of the probit and logit models are fairly close those of the probit model. These results suggest that very similar findings would be obtained using either the logit or the probit link function.

Based on the regression coefficients estimated in the main study (Table 5), the estimated distribution of (X, Z) and the estimates of the measurement error model parameters from the validation data, we back-calculated the regression coefficient to be expected in the corresponding model with no measurement error, using the attenuation matrix ([17], p. 53, Eq. (3.12); [43]-[45]). Finally, (4)-(5) were used to obtain the AUC and BS of each error-free probit model. These results are given in the last two columns of Table 6. By using the error-free measurements, the discrimination of the prediction model based on alcohol would likely be improved by 5.0% and the accuracy improved by 4.5%, while for the model based on $\alpha-$carotene, alcohol and Gail score, discrimination would not be improved but accuracy would likely be improved by 6.8%.

5. Summary

This paper examined the influence of measurement error on the performance of risk prediction models. We showed that measurement error in covariates, while not affecting calibration, can dramatically reduce the AUC and increase the BS. Thus, when it is possible to reduce the measurement error in predictors included in risk prediction models, the quality of the risk prediction could be improved substantially.

We also showed that the deterioration of the AUC and BS increases with measurement error. When the error-free determinants of the outcome are strongly positively correlated with the error-prone variable, and the error-free and error-prone covariates have similar correlations with the outcome, the measurement error reduces the AUC and increases the BS only slightly. The same is true when the error-free determinants of the outcome are strongly negatively correlated with the error-prone variable and the error-free and error-prone covariates have markedly different correlations with the outcome. The same is also true if auxiliary variables are available that are strongly correlated with the error-prone variable. We therefore recommend including in prediction models all easily obtained variables which are correlated with error-prone covariates, while avoiding overfitting by including too many parameters relative to the number of observations.

In addition, in the motivating example, we demonstrated how a validation study can be used to evaluate the impact of mismeasured covariates on risk prediction. This method can be used to evaluate the benefit of observing expensive but accurate data instead of the mismeasured data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97(18):1837–1847. [PubMed: 9603539]

2. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute. 1989; 81:1879–1886. [PubMed: 2593165]

3. Thompson IM, Ankerst DP, Chi C, Goodman PJ, Tangen CM, Lucia MS, Feng Z, Parnes HL, Coltman CA. Assessing Prostate Cancer Risk: Results from the Prostate Cancer Prevention Trial. Journal of the National Cancer Institute. 2006; 98(8):529–534. [PubMed: 16622122]
4. Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, Hennekens CH, Speizer FE. Reproducibility and validity of a semiquantitative food frequency questionnaire. American Journal of Epidemiology. 1985; 122(1):51–65. [PubMed: 4014201]
5. Chasan-Taber S, Rimm EB, Stampfer MJ, Spiegelman D, Colditz GA, Giovannucci E, Ascherio A, Willett WC. Reproducibility and validity of a self-administered physical activity questionnaire for male health professionals. Epidemiology. 1996; 7(1):81–86. [PubMed: 8664406]
6. Feskanich D, Rimm EB, Giovannucci EL, Colditz GA, Stampfer MJ, Litin LB, Willett WC. Reproducibility and validity of food intake measurements from a semiquantitative food frequency questionnaire. Journal of the American Dietetic Association. 1993; 93:790–796. [PubMed: 8320406]
7. Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. American Journal of Epidemiology. 1992; 135(10):1114–1126. [PubMed: 1632423]
8. Salvini S, Hunter DJ, Sampson L, Stampfer MJ, Colditz GA, Rosner BA, Willett WC. Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption. International Journal of Epidemiology. 1989; 18(4):858–867. [PubMed: 2621022]
9. Wolf A, Hunter DJ, Colditz GA, Manson JE, Stampfer MJ, Corsano K, Corsano KA, Rosner BA, Kriska A, Willett WC. Reproducibility and validity of a self-administered physical activity questionnaire. International Journal of Epidemiology. 1994; 23:991–999. [PubMed: 7860180]
10. Colditz GA, Stampfer MJ, Willett WC, Stason WB, Rosner BA, Hennekens CH, Speizer FE. Reproducibility and validity of self-reported menopausal status in a prospective cohort study. American Journal of Epidemiology. 1987; 126(2):319–325. [PubMed: 3605058]
11. Troy LM, Michels KB, Hunter DJ, Spiegelman D, Manson JE, Colditz GA, Stampfer MJ, Willett WC. Self-reported birthweight and history of having been breastfed among younger women: an assessment of validity. International Journal of Epidemiology. 1996; 25(1):122–127. [PubMed: 8666479]
12. Hunter DJ, Manson JE, Colditz GA, Chasan-Taber L, Troy L, Stampfer MJ, Speizer FE, Willett WC. Reproducibility of oral contraceptive histories and validity of hormone composition reported in a cohort of US women. Contraception. 1997; 56(6):373–378. [PubMed: 9494771]
13. Tomeo CA, Rich-Edwards JW, Michels KB, Berkey CS, Hunter DJ, Frazier AL, Willett WC, Buka SL. Reproducibility and validity of maternal recall of pregnancy-related events. Epidemiology. 1999; 10(6):774–777. [PubMed: 10535796]
14. Willett, WC. Reproducibility and Validity of Food-Frequency Questionnaires. Second Edition.. Oxford University Press; New York: 1998. Nutritional Epidemiology. Chapter 6.
15. Chasan-Taber S, Rimm EB, Stampfer MJ, Spiegelman D, Colditz GA, Giovannucci E, Ascherio A, Willett WC. Reproducibility and validity of a self-administered physical activity questionnaire for male health professionals. Epidemiology. 1996; 7(1):81–86. [PubMed: 8664406]
16. Fuller, WA. Measurement Error Models. John Wiley&Sons; New York: 1987.
17. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models. Chapman & Hall; London: 2006.
18. Buonaccorsi, JP. Measurement Error: Models, Methods, and Applications. Chapman & Hall (CRC Interdisciplinary Statistics); 2010.
19. Carroll RJ, Spiegelman CH, Lan KKG, Bailey KT, Abbott RD. On errors-in-variables for binary regression models. Biometrika. 1984; 71(1):19–25.
20. Carroll RJ, Wand MP. Semiparametric Estimation in Logistic Measurement Error Models. Journal of the Royal Statistical Society, Series B (Methodological). 1991; 53:573–585.
21. Li W, Mazumdar S, Arena VC, Sussman N. A resampling approach for adjustment in prediction models for covariate measurement error. Computer methods and programs in biomedicine. 2005; 77(3):199–207. [PubMed: 15721649]

22. Carroll RJ, Delaigle A, Hall P. Nonparametric prediction in measurement error models. Journal of the American Statistical Association. 2009; 104:993–1014. [PubMed: 20448838]

23. Colditz GA. The Nurses' Health Study: a cohort of US women followed since 1976. Journal of the American Medical Women's Association. 1994; 50(2):40–44.

24. Colditz GA, Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. American Journal of Epidemiology. 2000; 152(10): 950–964. [PubMed: 11092437]

25. Spiegelman D, Colditz GA, Hunter D, Hertzmark E. Validation of the Gail et al. model for predicting individual breast cancer risk. Journal of the National Cancer Institute. 86:600–607. [PubMed: 8145275]

26. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al. model of breast cancer risk prediction and implications of chemoprevention. Journal of the National Cancer Institute. 2001; 93:358–366. [PubMed: 11238697]

27. Garland M, Hunter DJ, Colditz GA, Spiegelman D, Manson JE, Stampfer MJ, Willett WC. Alcohol consumption in relation to breast cancer risk in a large cohort of US women. Cancer Epidemiology, Biomarkers and Prevention. 1999; 8:1017–1021.

28. Smith-Warner SA, Spiegelman D, Yaun S-S, Adami H-O, van den Brandt PA, Folsom A, Goldbohm RA, Graham S, Howe GR, Marshall JR, Miller AB, Potter JD, Speizer FE, Willett WC, Wolk A, Hunter DJ. Alcohol and breast cancer in women: a pooled analysis of cohort studies. JAMA. 1998; 279:535–540. [PubMed: 9480365]

29. Zhang X, Spiegelman D, Baglietto L, Bernstein L, Boggs DA, van den Brandt PA, Buring JE, Gapstur SM, Giles GG, Giovannucci E, Goodman G, Fraser G, Hankinson SE, Helzsouer KJ, Horn-Ross PL, Inoue M, Jung S, Khudyakov P, Larsson SC, Lof M, McCullough ML, Miller AB, Neuhouser ML, Palmer JR, Park Y, Robien K, Rohan TE, Ross JA, Schouten LJ, Shikany JM, Tsugane S, Visvanathan K, Weidarpass E, Wolk A, Willett WC, Zhang SM, Zeigler RG, Smith-Warner SA. Carotenoid intakes and risk of breast cancer defined by estrogen receptor and progesterone receptor status: a pooled analysis of 18 prospective cohort studies. The American Journal of Clinical Nutrituion. 2012; 95(3):713–725.

30. Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. American Journal of Epidemiology. 1992; 135(10):111426. discussion 1127-36.

31. Cox, DR.; Snell, EJ. Analysis of Binary Data. Chapman & Hall/CRC Monographs on Statistics & Applied Probability; 1989.

32. Greene, WH. Econometric Analysis. 3rd ed.. Prentice-Hall; Upper Saddle River, NJ: 1997.

33. Gill, J. Generalized Linear Models: A Unified Approach. Sage; Thousand Oaks, CA: 2001.

34. Long, JS. Regression Models for Categorical and Limited Dependent Variables. Sage; Thousand Oaks, CA: 1997.

35. Powers, DA.; Xie, Y. Statistical Methods for Categorical Data Analysis. Academic Press; San Diego: 2000.

36. Fahrmeir, L.; Tutz, G. Multivariate Statistical Modelling Based on Generalized Linear Models. 2nd ed.. Springer; New York: 2001.

37. Hardin, J.; Hilbe, J. Generalized Linear Models and Extensions. Stata Press; College Station, TX: 2001.

38. Hastie, T.; Tibshirani, R. Generalized Additive Models. Chapman and Hall; London: 1990.

39. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010; 21:128–138. [PubMed: 20010215]

40. Ikeda M, Ishigaki T, Yamauchi K. Relationship between Brier score and area under the binormal ROC curve. Computer Methods and Programs in Biomedicine. 2002; 67:187–194. [PubMed: 11853944]

41. Hernández-Orallo J, Flach P, Ferri C. A unified view of performance metrics: translating threshold choice into expected classification loss. The Journal of Machine Learning Research. 2012; 13(1): 2813–2869.

42. Falconer, DS.; Mackay, TFC.; Mackay, TFC.; MacKay, TF. Introduction to Quantitative Genetics. 4th edition. Longman Group Ltd. Assessment, and Prevention; New York: 1996.

43. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. American Journal of Epidemiology. 1990; 132(4):734–745. [PubMed: 2403114]

44. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. American Journal of Clinical Nutrition. 1997; 65:1179–1186.

45. Rosner B, Spiegelman D, Willett WC. Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Random Within-Person Measurement Error. American Journal of Epidemiology. 1992; 136(11):1400–1413. [PubMed: 1488967]

46. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev. 1950; 78:1–3.

47. Harrell Jr, F. Regression Modeling Strategies. Springer; NY: 2001.

48. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]

49. Wood, S. Package mgcv. R package version 1.7-29. 2014. http://cran.r-project.org/web/packages/mgcv/mgcv.pdf

50. Wood SN. mgcv: GAMs and generalized ridge regression for R. R news. 2001; 1(2):20–25.

51. Wood SN. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society, Series B. 2000; 62:413–428.

52. Sharma D, McGee D, Golam Kibria BM. Measures of explained variation and the base-rate problem for logistic regression. American Journal of Biostatistics. 2011; 2(1):11–19.

53. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. Journal of the National Cancer Institute. 2008; 100:1037–1041. [PubMed: 18612136]

54. Wacholder S, et al. Performance of common genetic variants in breast-cancer risk models. New England Journal of Medicine. 2010; 362:986–993. [PubMed: 20237344]

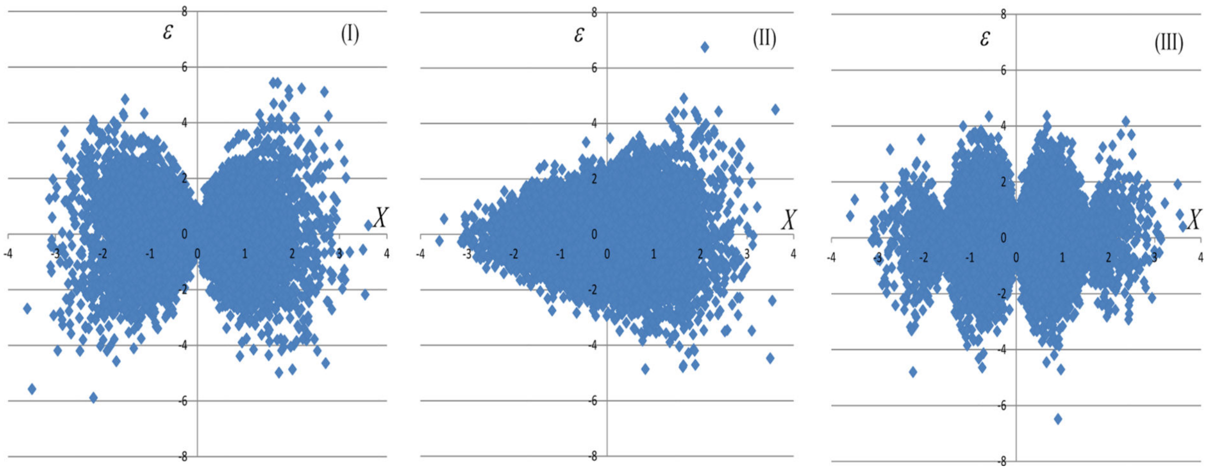

Figure 1.

Scatter plots of X versus ε under the heteroscedatic settings (I)-(III).

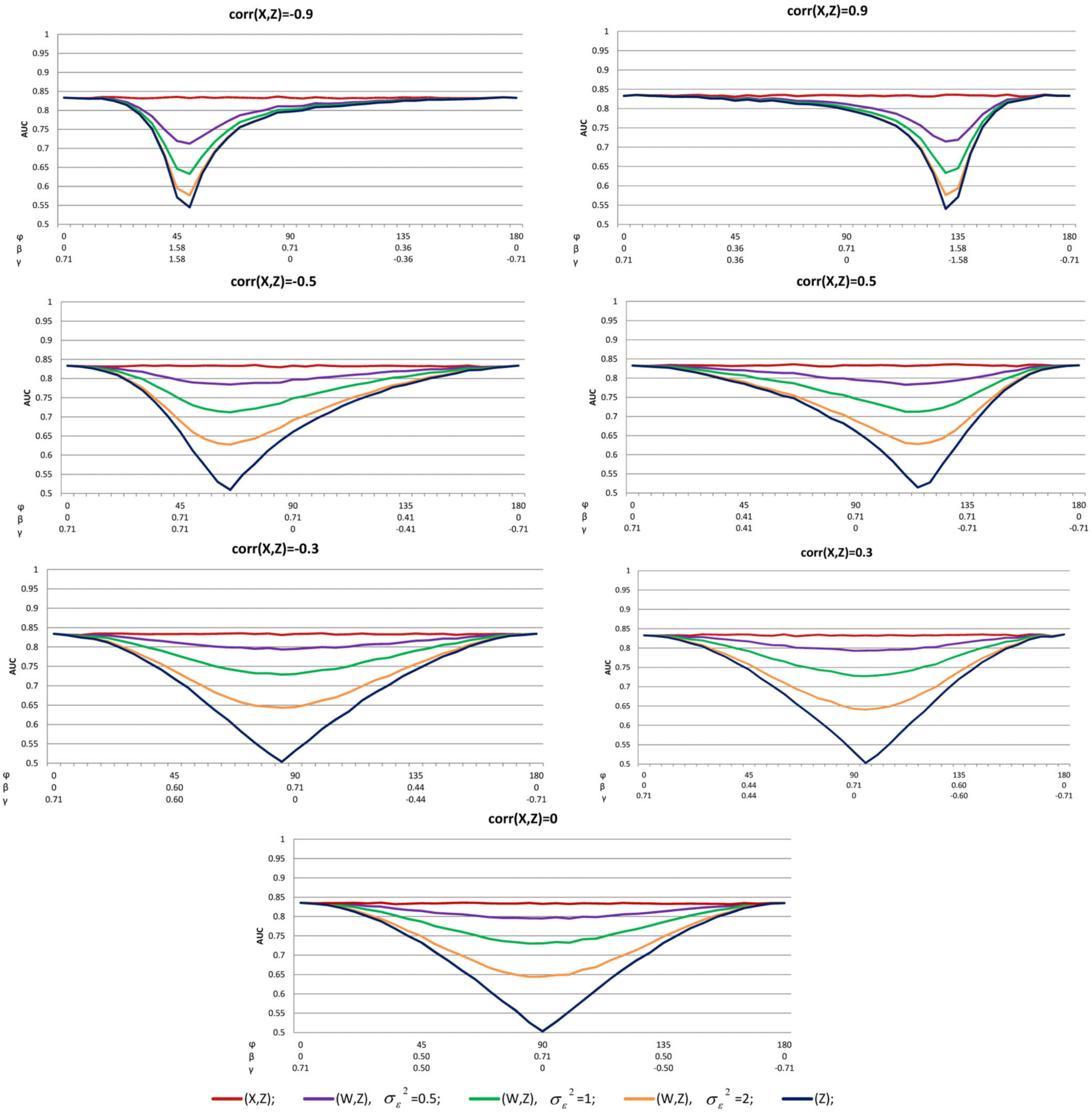

Figure 2.

AUCs of the homoscedastic error model, $R^2 = 0.5$, $p = 0.5$, and various values of $\rho_{X,Z} = corr(X, Z)$

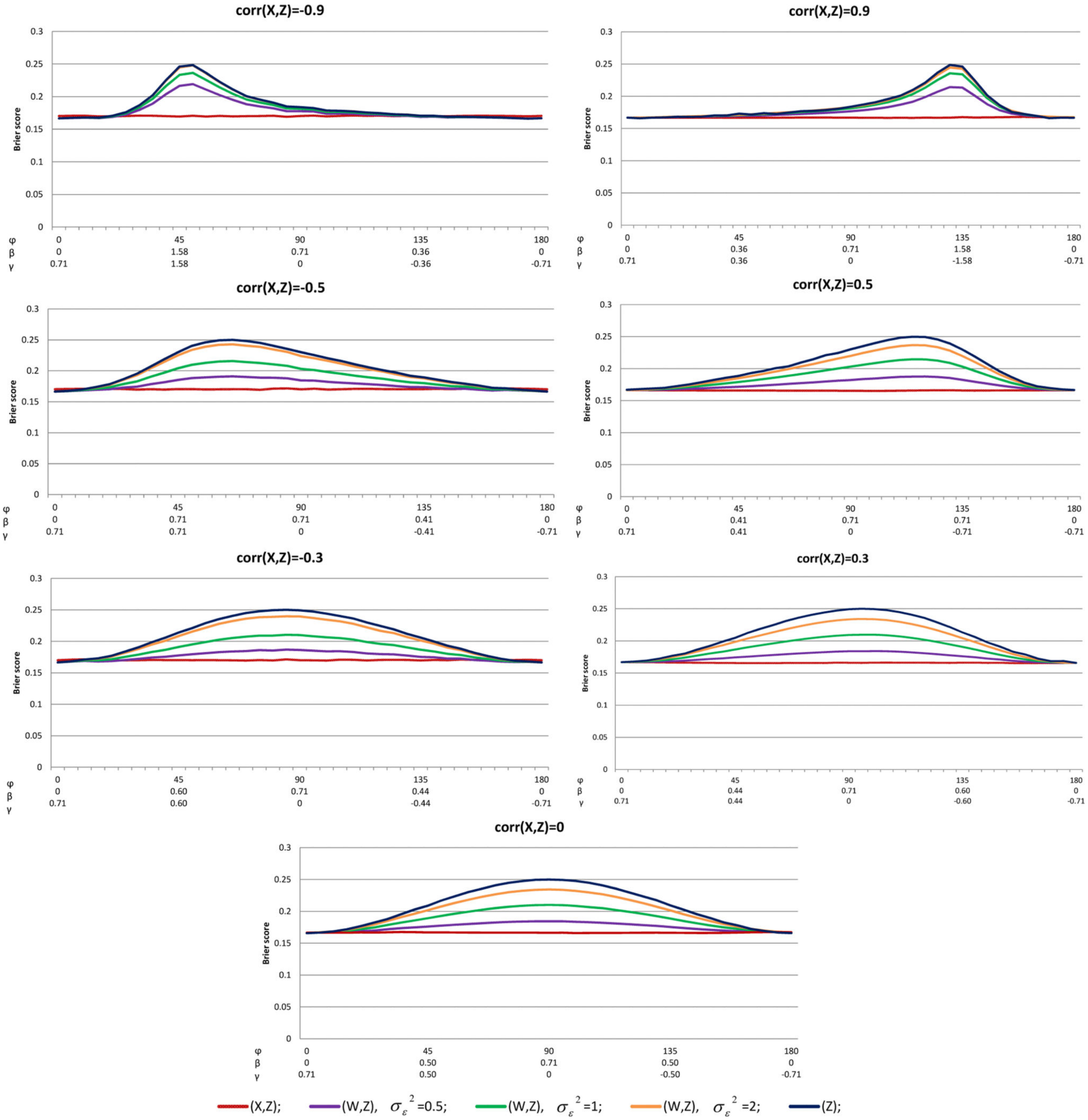

Figure 3.

BSs of the homoscedastic error model, $R^2 = 0.5$, $p = 0.5$, and various values of $\rho_{X,Z} = corr(X, Z)$

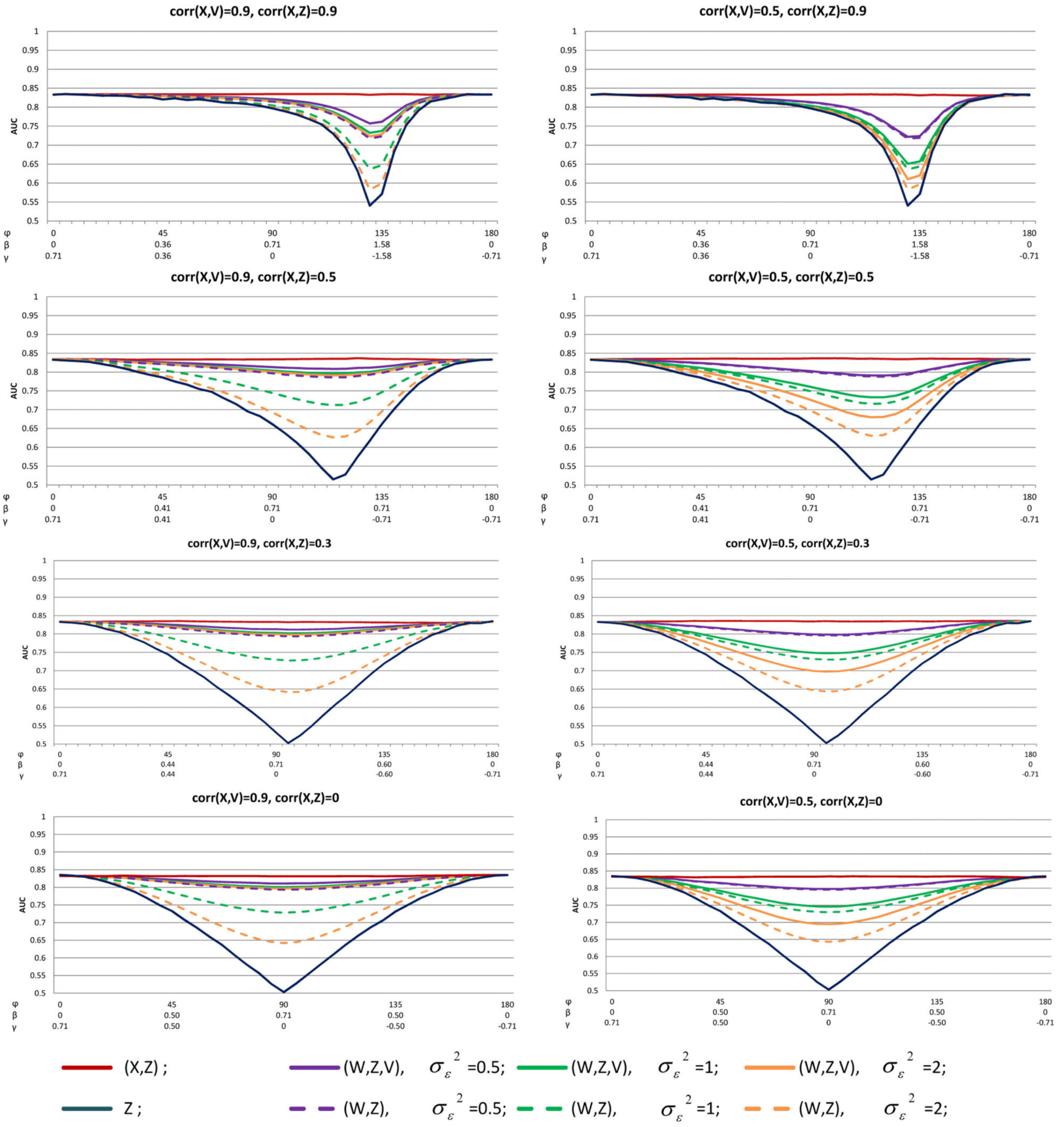

Figure 4.

The effect of AV on AUC: the homoscedastic error model, $R^2 = 0.5$, $p = 0.5$, and various values of $\rho_{X,Z} = corr(X, Z)$ and $\rho_{V,X} = corr(V, X)$

Table 1

AUC(X, Z) and BS(X, Z) as a function of R^2 and $p = \Pr(Y = 1)$

	AUC(X, Z)				***BS(X, Z)***			
	$p = \Pr(Y = 1)$				$p = \Pr(Y = 1)$			
R^2	**0.05**	**0.15**	**0.3**	**0.5**	**0.05**	**0.15**	**0.3**	**0.5**
0.02	0.588	0.568	0.567	0.565	0.046	0.127	0.207	0.247
0.04	0.624	0.600	0.592	0.591	0.047	0.124	0.204	0.244
0.06	0.652	0.626	0.613	0.609	0.046	0.125	0.203	0.241
0.08	0.668	0.641	0.630	0.626	0.047	0.125	0.200	0.238
0.10	0.692	0.661	0.651	0.641	0.046	0.122	0.197	0.235
0.12	0.705	0.681	0.663	0.655	0.047	0.120	0.195	0.231
0.14	0.721	0.695	0.676	0.671	0.045	0.119	0.193	0.228
0.16	0.734	0.704	0.690	0.681	0.045	0.118	0.189	0.225
0.18	0.749	0.720	0.701	0.696	0.045	0.116	0.188	0.220
0.20	0.762	0.734	0.712	0.706	0.045	0.115	0.185	0.217
0.30	0.819	0.781	0.761	0.754	0.042	0.109	0.172	0.201
0.40	0.863	0.824	0.803	0.796	0.042	0.102	0.158	0.184
0.50	0.900	0.862	0.841	0.836	0.038	0.093	0.143	0.166

Table 2

AUC(W, Z) under homoscedastic and heteroscedastic scenarios as a function of φ, $\sigma_{\varepsilon}^{2}=1$, $\rho_{X,Z}$ = 0, p = 0.5 *and AUC(X, Z)* = 0.834

				heteroscedastic					
			homoscedastic Probit	**scenario (I)**		**scenario (II)**		**scenario (III)**	
ϕ	β	γ		**GAM**	**Probit**	**GAM**	**Probit**	**GAM**	**Probit**
0°	0.000	0.707	0.834	0.833	0.832	0.834	0.834	0.833	0.833
15°	0.183	0.683	0.827	0.827	0.826	0.826	0.827	0.827	0.827
30°	0.354	0.612	0.810	0.809	0.812	0.811	0.812	0.810	0.810
45°	0.500	0.500	0.785	0.785	0.785	0.782	0.785	0.784	0.784
60°	0.612	0.354	0.758	0.758	0.758	0.753	0.756	0.761	0.761
75°	0.683	0.183	0.738	0.737	0.738	0.735	0.736	0.739	0.740
90°	0.707	0.000	0.729	0.727	0.725	0.729	0.729	0.730	0.730
105°	0.683	−0.183	0.739	0.736	0.740	0.737	0.738	0.738	0.738
120°	0.612	−0.354	0.759	0.759	0.762	0.756	0.759	0.761	0.762
135°	0.500	−0.500	0.784	0.785	0.785	0.782	0.785	0.786	0.786
150°	0.354	−0.612	0.810	0.810	0.809	0.806	0.808	0.809	0.810
165°	0.183	−0.683	0.827	0.828	0.827	0.824	0.825	0.826	0.826
180°	0.000	−0.707	0.834	0.834	0.832	0.833	0.833	0.832	0.832

Table 3

BS(W,Z) under homoscedastic and heteroscedastic scenarios as a function of φ, $\sigma_{\varepsilon}^{2}=1$, $\rho_{X,Z}=0$, $p=0.5$ *and* *BS (X,Z)* = 0.168

						heteroscedastic			
			homoscedastic Probit	**scenario (I)**		**scenario (II)**		**scenario (III)**	
ϕ	β	γ		**GAM**	**Probit**	**GAM**	**Probit**	**GAM**	**Probit**
0°	0.000	0.707	0.168	0.168	0.168	0.166	0.166	0.167	0.167
15°	0.183	0.683	0.170	0.170	0.170	0.170	0.170	0.170	0.170
30°	0.354	0.612	0.178	0.177	0.177	0.178	0.177	0.178	0.178
45°	0.500	0.500	0.189	0.187	0.187	0.191	0.189	0.189	0.189
60°	0.612	0.354	0.199	0.197	0.197	0.204	0.200	0.198	0.198
75°	0.683	0.183	0.207	0.203	0.203	0.211	0.208	0.206	0.206
90°	0.707	0.000	0.210	0.206	0.206	0.214	0.210	0.209	0.208
105°	0.683	0.183	0.207	0.204	0.204	0.210	0.207	0.206	0.206
120°	0.612	0.354	0.200	0.197	0.197	0.202	0.199	0.198	0.198
135°	0.500	0.500	0.189	0.187	0.187	0.191	0.189	0.188	0.188
150°	0.354	0.612	0.178	0.178	0.178	0.180	0.179	0.178	0.178
165°	0.183	0.683	0.169	0.170	0.170	0.171	0.171	0.170	0.170
180°	0.000	0.707	0.168	0.168	0.168	0.167	0.167	0.167	0.167

Table 4

AUCs and BS of probit model with binary regressors for various rates of misclassification q, as a function of $\rho_{X,Z}$

			AUC(W,Z̃)˜						***BS(W,Z̃)˜***			
			q						***q***			
$\rho_{X,Z}$	**AUC(*Z*)˜**	***AUC(X,Z̃)˜***	**0.05**	**0.15**	**0.25**	**0.5**	***BS(Z)˜***	***BS(X,Z̃)˜***	**0.05**	**0.15**	**0.25**	**0.5**
−0.9	0.545	0.611	0.600	0.581	0.564	0.545	0.148	0.145	0.146	0.147	0.147	0.148
−0.7	0.616	0.725	0.708	0.675	0.648	0.616	0.163	0.149	0.152	0.158	0.161	0.163
−0.5	0.671	0.785	0.768	0.732	0.705	0.671	0.169	0.146	0.150	0.159	0.164	0.169
−0.3	0.717	0.826	0.809	0.774	0.749	0.717	0.169	0.139	0.145	0.155	0.162	0.169
0	0.776	0.865	0.850	0.822	0.800	0.776	0.162	0.130	0.136	0.146	0.154	0.162
0.3	0.828	0.892	0.881	0.860	0.845	0.828	0.148	0.120	0.125	0.135	0.141	0.148
0.5	0.862	0.907	0.898	0.884	0.874	0.862	0.135	0.113	0.118	0.125	0.130	0.135
0.7	0.895	0.919	0.913	0.905	0.900	0.895	0.120	0.106	0.110	0.114	0.117	0.120
0.9	0.925	0.931	0.929	0.927	0.926	0.925	0.102	0.099	0.100	0.101	0.102	0.102

Table 5

NHS data: descriptive statistics from validation study and parameter estimates of probit models for breast cancer outcome from main study

Validation study	α-carotene	Alcohol	Gail Score
Mean of error-free measure	583.3	9.2	−2.3
Variance of error-free measure	142280.3	138.4	0.023
Intercept and slope from measurement error model (*c*, *d*)	(507.1,0.844)	(1.393,0.645)	
Error variance - σ_{ε}^{2}	246909.6	53.7	-
Correlation between error-prone variable and its surrogate	0.540	0.721	-

Main study	Regression coefficients (standard error)			
Probit regression model	**intercept**	**α-carotene (mg/day)**	**Alcohol (g/day)**	**Gail Score**
α-carotene	−1.693(0.013)	1.37e-05(1.37e-05)	-	-
α-carotene+Gail Score	0.322(0.124)	−9.86e-06(1.43e-05)	-	0.888(0.054)
Alcohol	−1.702(0.010)	-	0.003(0.001)	-
Alcohol+Gail Score	0.278(0.122)	-	0.003(0.001)	0.879(0.054)
α-carotene+Alcohol	−1.715(0.014)	1.67e-05(1.37e-05)	0.003(0.001)	-
α-carotene+Alcohol+Gail Score	0.289(0.124)	−6.99e-06(1.43e-05)	0.003(0.001)	0.882(0.055)

Table 6

NHS main study: AUC, BS and −2 log-likelihood (−2LL) values for breast cancer risk prediction of the surrogate model and the true model

	Surrogate Probit Model			Surrogate Logit Model			Surrogate GAM Model		Error-Free Probit Model	
Regression Model	**AUC**	**BS**	**−2LL**	**AUC**	**BS**	**−2LL**	**AUC**	**BS**	**AUC**	**BS**
α-carotene	0.509	0.044	12340.4	0.509	0.044	12340.4	0.500	0.044	0.505	0.046
α-carotene+GS	0.586	0.044	12210.0	0.586	0.044	12209.5	0.585	0.044	0.583	0.043
Alcohol	0.500	0.044	12335.5	0.500	0.044	12335.5	0.521	0.044	0.525	0.046
alcohol+GS	0.588	0.044	12205.7	0.588	0.044	12205	0.588	0.044	0.585	0.041
α-carotene+alcohol	0.522	0.044	12333.9	0.522	0.044	12333.8	0.519	0.044	0.530	0.051
α-carotene+alcohol+GS	0.588	0.044	12205.7	0.588	0.044	12205	0.588	0.044	0.586	0.041