STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:987–998 (DOI: 10.1002/sim.1662)

Estimation of magnitude in gene–environment interactions in the presence of measurement error

M. Y. Wong[1], N. E. Day[2,*,†], J. A. Luan[3] and N. J. Wareham[3]

[1]*Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China*
[2]*Strangeways Research Laboratory, Institute of Public Health, University of Cambridge, Cambridge CB1 8RN, U.K.*
[3]*Department of Public Health and Primary Care, Institute of Public Health, University of Cambridge, Cambridge CB2 2SR, U.K.*

SUMMARY

The design of studies aimed at identifying the interaction between genetic and environmental determinants in disease risk is attracting increased attention. In this paper, we study the effect of errors on measuring exposures and the effect of assessing genotype at one locus on the association of a continuous outcome measure with continuous exposures and genotype. The estimation of misclassification errors in assessing genotypes from a separate study is proposed.
If the exposure measurement error is substantial, then even relatively small errors in genotyping within limits that are often quoted can have an appreciable effect on interaction estimates. We, thus, consider a method for correcting the measurement errors when the interaction between the exposures and the genetic factor is significant. Finally, we present an epidemiological example to demonstrate the importance of correcting measurement errors. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: genotype; environmental exposure; gene–environment interaction; measurement error model

1. INTRODUCTION

Studies for identifying interaction between genetic and environmental factors in disease risk have become of increasing importance. We have previously described power calculations for situations in which both the exposure and the outcome are continuously distributed [1] and we have extended those calculations to take account of measurement error [2]. In this paper, we concentrate on the estimation of gene–environment interaction when both genotype and the exposure are observed with error.

*Correspondence to: Professor N. E. Day, Strangeways Research Laboratory, Institute of Puplic Health, University of Cambridge, Cambridge CB1 8RN, U.K.
†E-mail: nick.day@srl.cam.ac.uk

Contract/grant sponsor: The Royal Society

Received July 2002
Accepted July 2003

Our model for gene–environment interactions, defining the response y, such as blood pressure, for genotype j, $j=1,2,3$, with a single continuous exposure T, is

$$y_j = \alpha_j + \beta_j T + \text{error}$$

where α_j is the main effect for genotype j and β_j is the regression coefficient of y on T for individuals with genotype j. The error is assumed to be random with mean zero and variance σ_y^2 for all j. The differences between the slopes in the groups apart from the last group are defined as interaction terms. The results are generalized to multivariate exposures, which can be found on URL address: http://www.math.ust.hk/∼mamywong/paper/sim_2003.pdf.

We focus on the estimation of regression coefficients in our model when there are errors in measuring exposure and in assessing genotype. We also assess the precision with which the effect of the measurement error and the misclassification error on the crude estimates of the regression coefficients has been adjusted. Finally, we present an epidemiological example to demonstrate the extent to which ignoring uncertainty in the exposure and genotyping assessment can distort the conclusion.

2. METHOD

2.1. Measurement error model and misclassification rate

Because of measurement error, instead of the true covariate, T, we observe its respective surrogate, R. We assume that error is non-differential with regard to the outcome variable, y, that is, R contributes no information about y beyond what is available in T. R is related to T by an additive error model as $R = T + \varepsilon_R$ with $\mathrm{E}(\varepsilon_R) = 0$ and $\mathrm{Var}(\varepsilon_R) = \sigma_R^2$. The true latent variable, T, is assumed to have mean μ and variance σ_T^2.

Suppose that there are two different alleles in a locus, a and A, where a and A are rare and common alleles, respectively. We consider a general situation for a polymorphism with frequency p of the common allele. Assume the polymorphism is in the Hardy–Weinberg equilibrium [3]. Then the genotype frequencies of aa, Aa and AA are $p_1 = (1-p)^2$, $p_2 = 2p(1-p)$, $p_3 = p^2$, respectively. We assume that the misclassification of each allele is independent of the misclassification of another. Let the probabilities of misclassification of a and A be P_a and P_A, respectively. The observed genotype frequency of the common allele is then equal to $p' = p(1 - P_A) + (1 - p)P_a$. The observed frequencies of aa, Aa and AA, thus, become $p'_1 = (1 - p')^2$, $p'_2 = 2p'(1 - p')$ and $p'_3 = p'^2$.

2.2. Estimates of regression coefficients

It is known that crude estimates $\hat{\alpha}_j$ and $\hat{\beta}_j$ for α_j and β_j, for $j=1,2,3$, obtained by ignoring the errors on measuring exposure and assessing genotype are biased. For all formulas in Section 2, we assign three genotypes aa, Aa and AA to be genotypes 1, 2 and 3, respectively.

For simplicity, we assume that μ is equal to zero here. It can be shown that

$$\begin{pmatrix} \mathrm{E}(\hat{\alpha}_1) \\ \mathrm{E}(\hat{\alpha}_2) \\ \mathrm{E}(\hat{\alpha}_3) \\ \mathrm{E}(\hat{\beta}_1) \\ \mathrm{E}(\hat{\beta}_2) \\ \mathrm{E}(\hat{\beta}_3) \end{pmatrix} = \begin{pmatrix} \Delta & 0 \\ 0 & (\sigma_T^2 + \sigma_R^2)\Delta \end{pmatrix}^{-1} \begin{pmatrix} \Delta_e & 0 \\ 0 & \sigma_T^2 \Delta_e \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

where

$$\Delta = \begin{pmatrix} p_1' & 0 & 0 \\ 0 & p_2' & 0 \\ 0 & 0 & p_3' \end{pmatrix}$$

and

$$\Delta_e = \begin{pmatrix} p_1(1-P_a)^2 & p_2(1-P_a)P_A & p_3 P_A^2 \\ 2p_1(1-P_a)P_a & p_2((1-P_a)(1-P_A)+P_aP_A) & 2p_3P_A(1-P_A) \\ p_1P_a^2 & p_2P_a(1-P_A) & p_3(1-P_A)^2 \end{pmatrix}$$

The (i,j)th element of Δ_e is the joint probability that an individual has been assessed to have genotype i but he has, in fact, the genotype j. The expressions for a non-zero μ can be found on URL address: http://www.math.ust.hk/∼mamywong/paper/sim_2003.pdf.

In order to obtain unbiased estimates of the association between T and y, the crude biased estimates need correction. The adjusted estimates for $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ and β_3 are equal to

$$\begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \tilde{\alpha}_3 \end{pmatrix} = \Delta_e^{-1}\Delta \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{pmatrix}$$

and

$$\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{\beta}_3 \end{pmatrix} = \frac{\sigma_T^2 + \sigma_R^2}{\sigma_T^2} \Delta_e^{-1}\Delta \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$$

If there is no misclassification error on assessing genotype, the estimates of the regression coefficients α_1, α_2 and α_3 are affected by errors on measuring exposure only if μ is not equal to zero. It is well known that the observed outcome–exposure relationship is on average weaker than the true outcome–exposure relationship because of the non-differential measurement error

in the exposure. The crude estimate is attenuated towards zero. An asymptotical unbiased estimate can be obtained from multiplying the crude estimate by a correction factor [4, 5].

It should be noted that the model used in this paper can be easily extended to the situation with a categorical variable with any number of classes. In the general situation, Δ and Δ_e are $k \times k$ matrices where k is the number of classes for the categorical variable. Δ is a diagonal matrix with the ith diagonal element equal to the proportion of subjects in ith class. The (i, j)th element in Δ_e is the joint probability of subjects being assigned to ith class but in fact belonging to jth class.

2.3. Validation study

In order to obtain unbiased estimates of the association between the exposures and the outcome, we need to estimate the relevant correction factor. The sub-study required to estimate the unknown parameters in the correction factor is called a validation study [5]. In our previous papers [4, 5], we have discussed how to estimate the unknown parameters in the measurement model in exposures for two types of validation studies. In this paper, we discuss only the estimation of the misclassification rate in assessing genotypes.

Misclassification is a common problem in genetic epidemiology [3]. Duffy *et al.* [6] described the method for analysing a case-control study of breast cancer risk and such lifestyle attributes as diet, smoking and alcohol consumption, in which misclassification was anticipated and, consequently, repeatability data were collected. Duffy *et al.* [7] considered the difference in precision of risk estimate when the misclassification of a binary risk factor was corrected by two strategies. The first one was to estimate the required correction to the risk estimates from a validation study, independent of the epidemiological study under consideration (external validation) [8]. The second was to measure the quantities subject to error repeatedly within the epidemiological study (internal repeat measurement) [9]. In this paper, we consider the situation which genotypes are defined by two alleles of a single polymorphism, that is three genotypes, *aa*, *Aa* and *AA*.

Since a 'gold standard' usually does not exist, for estimating the misclassification rates in assessing genotypes, each individual is genotyped twice in our validation study. Suppose we have data in the form given below:

	2nd determination			
1st determination	*aa*	*Aa*	*AA*	
aa	n_{11}	n_{12}	n_{13}	$n_{1.}$,
Aa	n_{21}	n_{22}	n_{23}	$n_{2.}$,
AA	n_{31}	n_{32}	n_{33}	$n_{3.}$,
	$n_{.1}$	$n_{.2}$	$n_{.3}$	

Assuming the independence of repeat determinations conditional on the true state, we have

$$\Pr(\text{observing '}a\text{' in both determinations}) = P_{aa} = pP_A^2 + (1 - p)(1 - P_a)^2$$

$$\Pr(\text{observing '}A\text{' in both determinations}) = P_{AA} = p(1 - P_A)^2 + (1 - p)P_a^2$$

$$\Pr(\text{observing } `a' \text{ in one determination and } `A' \text{ in the other}) = P_{Aa}$$

$$= pP_A(1-P_A) + (1-p)P_a(1-P_a)$$

It is noted that $P_{Aa} = (1 - P_{aa} - P_{AA})/2$. Thus, assuming the independence of two alleles of a gene, the probability of observing each combination of two determinations is given as follows:

	2nd determination		
1st determination	*aa*	*Aa*	*AA*
aa	P_{aa}^2	$2P_{aa}P_{Aa}$	P_{Aa}^2
Aa	$2P_{aa}P_{Aa}$	$2(P_{aa}P_{AA} + P_{Aa}^2)$	$2P_{Aa}P_{AA}$
AA	P_{Aa}^2	$2P_{Aa}P_{AA}$	P_{AA}^2

The likelihood function is thus proportional to

$$L \propto P_{aa}^{n_{aa}} P_{AA}^{n_{AA}} (1 - P_{aa} - P_{AA})^{n_{Aa}} \{P_{aa}P_{AA} + (1 - P_{aa} - P_{AA})^2/4\}^{n_{22}}$$

where $n_{aa} = 2n_{11} + n_{12} + n_{21}$, $n_{Aa} = n_{12} + 2n_{13} + n_{21} + n_{23} + 2n_{31} + n_{32}$ and $n_{AA} = n_{23} + n_{32} + 2n_{33}$. The numerical values of P_{aa} and P_{AA} can be obtained by solving $\partial \log L/\partial P_{aa} = 0$ and $\partial \log L/\partial P_{AA} = 0$. By substituting P_{aa} and P_{AA} with $pP_A^2 + (1-p)(1-P_a)^2$ and $p(1-P_A)^2 + (1-p)P_a^2$, respectively, we have two equations in terms of three quantities of p, P_a and P_A. We hence need additional information. In this paper, we assume that $P_a = P_A = P_m$. Under this assumption, it is called random error model [10–12]. On some occasions, p is known externally and then P_a and P_A can be estimated separately.

2.4. Precision

Since the adjusted estimates can be obtained by multiplying the crude estimates by a correction factor and the unknown parameters in the correction factor, including error in measuring exposure, the correlation between true and observed exposures, the frequency of the common allele and the probabilities of misclassification in assessing genotypes, are estimated by a validation study, their variances and covariances thus depend on the variation from both the main and validation studies. Denote ϕ_i, $i = 1, \ldots, l$, as the regression coefficients in the model. The adjusted estimates, $\tilde{\phi}_1, \ldots, \tilde{\phi}_l$, are thus linear combinations of the crude estimates, $\hat{\phi}_1, \ldots, \hat{\phi}_l$, that is $\tilde{\phi}_i = \sum_{j=1}^{l} a_{ij}\hat{\phi}_j$, where a_{ij} is a function of statistics from the validation study. The covariance of $\tilde{\phi}_i$ and $\tilde{\phi}_j$ is thus approximately equal to, for $i, j = 1, \ldots, k$,

$$\mathrm{Cov}(\tilde{\phi}_i, \tilde{\phi}_j) \approx \sum_{s=1}^{k}\sum_{t=1}^{k} (\mathrm{Cov}(a_{is}, a_{jt})\mathrm{E}(\hat{\phi}_s, \hat{\phi}_t) + \mathrm{Cov}(\hat{\phi}_s, \hat{\phi}_t)\mathrm{E}(a_{is})\mathrm{E}(a_{jt}))$$

In this section, we discuss the variation due to the estimation of unknown parameters in the measurement error in exposure and the misclassification rate in genotype assessment only.

For $s = 1, 2, 3$,

$$\mathrm{Var}(\tilde{\alpha}_s) = \alpha^{\mathrm{T}} D_{ss} \alpha$$

$$\mathrm{Var}(\tilde{\beta}_s) = \beta^{\mathrm{T}} (\mathrm{Var}(\overline{\mathrm{CF}}) r_s r_t^{\mathrm{T}} + \overline{\mathrm{CF}}^2 D_{ss}) \beta$$

where $\overline{CF}$ is an estimate of the univariate correction factor for the exposure and is equal to $(\hat{\sigma}_T{}^2 + \hat{\sigma}_R{}^2)/\hat{\sigma}_T{}^2$; r_i^{T} is the ith row of $\hat{\Delta}_e{}^{-1}\hat{\Delta}$; D_{ij}, a 3×3 matrix, is the covariance matrix of $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$.

Since elements in $\hat{\Delta}_e{}^{-1}\hat{\Delta}$ can be expressed as functions of the maximum likelihood estimates of p, P_a and P_A, the variance and covariance matrix of the elements in $\hat{\Delta}_e{}^{-1}\hat{\Delta}$ can be obtained by Taylor expansion. The variance and covariance matrix of the maximum likelihood estimates of p, P_a and P_A can be found on URL address: http://www.math.ust.hk/~mamywong/paper/sim_2003.pdf.

3. RESULTS

3.1. Bias resulting from the measurement error

For the purpose of illustrating the results in Section 2, we consider a dominant model, that is, carriers of the rare allele versus homozygotes for the common allele. In this case, Δ and Δ_e become 2×2 matrices with

$$\Delta = \begin{pmatrix} 1 - p'^2 & 0 \\ 0 & p'^2 \end{pmatrix}$$

and

$$\Delta_e = \begin{pmatrix} p_1(1 - P_a^2) + p_2(1 - P_a(1 - P_A)) & p_3(1 - (1 - P_A)^2) \\ p_1 P_a^2 + p_2 P_a (1 - P_A) & p_3 (1 - P_A)^2 \end{pmatrix}$$

where $p' = p(1 - P_A) + (1 - p)P_a$, $p_1 = (1 - p)^2$, $p_2 = 2p(1 - p)$ and $p_3 = p^2$ with p the frequency of the common allele and P_a and P_A the probabilities of misclassification of a and A, respectively.

From the expressions of expectations of crude estimates in Section 2.2, it is noted that the values of slopes and the correlation coefficient between the observed and the true exposures, ρ_x, do not affect the bias of the crude estimate for the intercept if μ is set equal to zero. The biases of the crude estimates for slopes do not depend on the values of intercept and μ. When the two intercepts are equal, their corresponding crude estimates are unbiased. When the two slopes are equal, the expectations of their corresponding crude estimates are equal to their true values divided by the correction factor (that is, multiplied by ρ_x^2).

In Table I, we give the expectations of the crude estimates of the slopes in the regression lines of the outcome of the exposure for the group of individuals with at least one rare allele and the group of individuals with two common alleles and their ratio. The expectations of the crude estimates of the intercepts are equal to those of the slopes when $\rho_x = 1$. The expectations

Table I. The expectations of the crude estimates of the slopes in the regression lines of the outcome of the exposure for the group of individuals with at least one rare allele and the group of individuals with two common alleles and their ratio. P_A and P_a are set equal to P_m. β_2 is set equal to one.

			$\rho_x = 0.4$		$\rho_x = 0.8$		$\rho_x = 1$		
β_1/β_2	p	P_m	$E(\hat{\beta}_1)$	$E(\hat{\beta}_2)$	$E(\hat{\beta}_1)$	$E(\hat{\beta}_2)$	$E(\hat{\beta}_1)$	$E(\hat{\beta}_2)$	$E(\hat{\beta}_1/\hat{\beta}_2)$
1.5	0.8	0.01	0.2372	0.1604	0.9490	0.6416	1.4828	1.0025	1.4791
		0.05	0.2277	0.1621	0.9110	0.6483	1.4234	1.0129	1.4052
		0.1	0.2185	0.1643	0.8740	0.6571	1.3656	1.0267	1.3301
	0.95	0.01	0.2275	0.1601	0.9098	0.6403	1.4216	1.0005	1.4208
		0.05	0.2011	0.1604	0.8044	0.6418	1.2569	1.0028	1.2534
		0.1	0.1873	0.1609	0.7493	0.6437	1.1708	1.0058	1.1640
3	0.8	0.01	0.4690	0.1615	1.8759	0.6464	2.9311	1.0101	2.9019
		0.05	0.4310	0.1683	1.7238	0.6730	2.6934	1.0516	2.5613
		0.1	0.3940	0.1771	1.5760	0.7083	2.4624	1.1067	2.2251
	0.95	0.01	0.4298	0.1604	1.7193	0.6414	2.6863	1.0021	2.6807
		0.05	0.3244	0.1617	1.2976	0.6471	2.0276	1.0110	2.0054
		0.1	0.2693	0.1637	1.0771	0.6548	1.6830	1.0232	1.6449

of the slopes for other values of β_2 are obtained by multiplying the expectations in Table I by β_2 for the same ratio of β_1 and β_2. The expectation of the ratio of two slope estimates does not depend on the value of β_2.

The data in Table I indicate that the estimate for the smaller intercept is over-estimated and that for the larger intercept is under-estimated. The bias for the slope increases as the frequency of the common allele increases and as the correlation coefficient between the true and observed exposures decreases. The bias of the ratio of two slope estimates does not depend on the correlation coefficient between the true and observed exposures. It increases with increments in the frequency of the common allele, the probability of misclassification and the true ratio. The ratio is under-estimated if it is greater than 1. Otherwise, the ratio is over-estimated.

In Table II, we give the standard error of the ratio of two slope estimates in the regression lines of the outcome on the exposure for the group of individuals with at least one rare allele and the group of individuals with two common alleles. Only the variation from the estimation of unknown parameters in the measurement error model is considered when calculating the standard error in Table II. The variation from the main study of modelling the gene–environment interaction is ignored. We assume that the validation study for estimating the reliability coefficient for the exposure has 100 subjects, each with two repeated measures. For estimating the genotyping error, 1000 subjects are assumed to be genotyped twice.

Although the ratio of the two slopes does not depend on the correlation coefficient between the true and observed exposures, its standard error does. This standard error depends on the frequency of the common allele, the probability of misclassification, the true ratio and the correlation coefficient between the true and observed exposures, ρ_x. The standard error of the ratio of the two slopes is inversely proportional to the value of β_2. As ρ_x changes from

Table II. The standard error of the ratio of two slope estimates in the regression lines of the outcome on the exposure for the group of individuals with at least one rare allele and the group of individuals with two common alleles. P_A and P_a are set equal to P_m. β_2 is set equal to one.

β_1/β_2	p	P_m	$\rho_x = 0.4$	$\rho_x = 0.8$	$\rho_x = 1$
1.5	0.8	0.01	0.0085	0.0042	0.0034
		0.05	0.0184	0.0092	0.0073
		0.1	0.0254	0.0127	0.0102
	0.95	0.01	0.0323	0.0161	0.0129
		0.05	0.0463	0.0232	0.0185
		0.1	0.0493	0.0247	0.0197
3	0.8	0.01	0.0388	0.0194	0.0155
		0.05	0.0778	0.0389	0.0311
		0.1	0.0997	0.0499	0.0400
	0.95	0.01	0.1294	0.0647	0.0517
		0.05	0.1829	0.0915	0.0732
		0.1	0.1915	0.0957	0.0766

1 to 0.4, the standard error increases by roughly 2.5-fold for each situation considered. This is an increase in variance of 6.25, implying the need for a sample size 6.25 times larger to achieve the same precision from the validation study. When the misclassification rate increases by 10-fold (from 1 to 10 per cent), the standard error increases by roughly 3-fold when the probability of the common allele is 0.8. The increment decreases as the probability of the common allele increases.

3.2. An epidemiological example

To illustrate the practical application of this correction method, we have used the data from the Isle of Ely Study [13, 14] to examine the interaction of habitual energy expenditure with the polymorphism G-62A within the G-protein-coupled receptor GPR10 in the determination of blood pressure. We seek to determine whether physical activity, which is known to affect an individual's blood pressure [15], interacts with the genotype at the GPR10 locus to affect systolic blood pressure (SBP) and diastolic blood pressure (DBP) before and after correcting errors on measuring habitual energy expenditure and in assessing genotypes. Wareham [16] developed an objective method—the heart rate monitoring with individual calibration—which produces a quantitative estimate of energy expenditure over 4 days. This estimate is expressed as the physical activity level (PAL), the ratio of total energy expenditure to basal metabolic rate. In this population-based study of 821 adults, all subjects had their PAL measured on a single occasion and their G-62A polymorphism allele genotyped (G: common allele, A: rare allele).

The reliability coefficient for PAL is computed from a random group of 197 subjects in the cohort from whom PAL measurements were taken on a further three occasions. This coefficient is estimated to be 0.5653.

Table III. Association of centred and standardized energy expenditure (PAL) with diastolic blood pressure. Results of logistic regression adjusted by sex, centred age and centred BMI with correction for measurement error of PAL and misclassification error in assessing genotype. The value inside the parentheses is the standard error of the estimate.

		Intercept		
Estimate	Genotype	Female	Male	Slope
Crude	G/A & A/A	74.29 (0.58)	80.12 (0.63)	−0.61 (0.50)
	G/G	73.18 (0.52)	79.01 (0.59)	−2.03 (0.45)
Corrected	G/A & A/A	74.31 (0.72)	80.15 (0.74)	−1.02 (0.90)
	G/G	73.17 (0.61)	79.01 (0.65)	−3.61 (0.83)

In a validation study for estimating the frequency of the common allele and the misclassification, 1027 subjects were genotyped twice. The data are given as follows:

	2nd determination			
1st determination	G/G	G/A	G/A	
G/G	573	10	0	583
G/A	15	363	1	379
A/A	0	4	61	65
	588	377	62	1027

Since the frequency of the common allele is unknown in this study, we assume two misclassification rates equal. Using the method of maximum likelihood, the frequency of the common allele and the common misclassification rate are estimated to be 0.758 and 0.0074, respectively.

We combine rare allele carriers into a group when we perform the statistical analysis. We also centre and standardize PAL to stabilize crude estimates. We only consider the effect of the interaction of PAL and G-62A on diastolic blood pressure in this paper.

Table III gives the results of the multiple regression analysis relating diastolic blood pressure to PAL, taking sex, age and body mass index (BMI) into account. The slopes for both groups after correcting the measurement error would be equal to multiplying the corresponding crude estimates by the correction factor, which equals the inverse of the reliability coefficient and is estimated to be 1.769, if there were no error in assessing the genotype. With a small genotype misclassification rate estimated to be 0.0074, the correction factors for the two groups are, thus, 1.68 and 1.78, respectively. The estimates of the interaction term (i.e. $\beta_1 - \beta_2$) before and after correction are 1.43 (s.e. = 0.66) and 2.59 (s.e. = 1.20), respectively. The test for the equality of the slopes based on the crude estimation and the estimation after correction are thus significant with the p-values being 0.03 and 0.031, respectively.

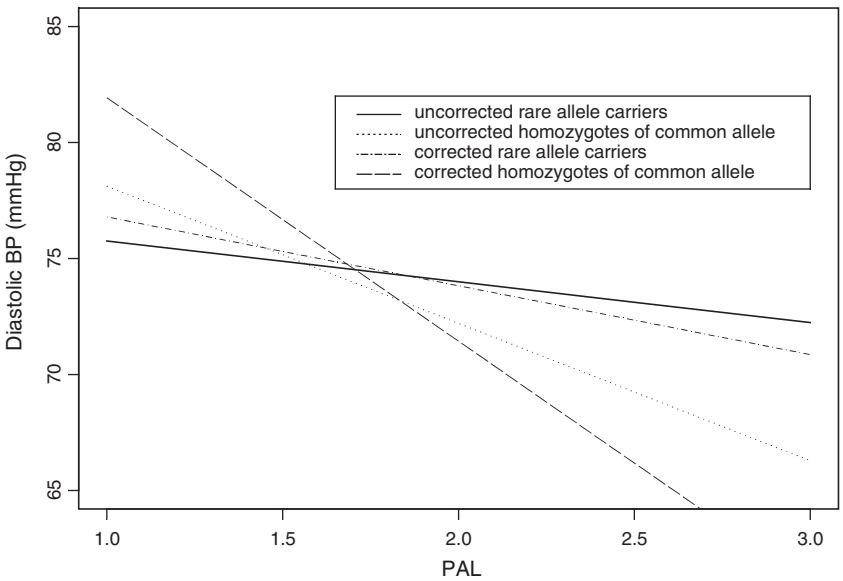

Figure 1. Diastolic blood pressure for females at mean age (= 53.81) and mean BMI (= 26.38) versus different values of PAL.

Since the slopes are significantly different, the effect of energy expenditure on diastolic blood pressure is different for different groups. It is noted from Table III that the relationship between blood pressure and PAL is not significant for the group of rare allele carriers, whether before or after correction of errors. The relationship is highly significant for the homozygotes with the common allele. This means that the influence of habitual energy expenditure on blood pressure is significant for the homozygotes of the common allele.

Figure 1 displays graphically the crude and corrected regressions shown in Table III. It is clear that the main impact of the regression correction is to accentuate the reduction in blood pressure associated with high levels of physical activity.

4. DISCUSSION

Measurement error is an unavoidable aspect of observational epidemiology. The effect of measurement error on the observed relationship between exposure and disease outcome has been extensively studied, as has the effect of genotyping errors in linkage studies [12]. However, even with greater emphasis being put on the combined effect of genetic and environmental factors, little attention has been paid to the effect of measurement error on the estimation of such interactions. In this situation, measurement error, either in genotyping or in exposure measurement, may seriously distort the observed interaction. In the example we have given, the absolute difference between two genotypes in the slope relating energy expenditure to blood pressure is substantially under-estimated if no allowance is made for measurement error, as can be seen in Figure 1. The implication of the corrected relationship is rather dif-

ferent from that of the uncorrected relationship. Among homozygotes for the common allele, the effect of energy expenditure on blood pressure can be seen to be of much greater public health significance once one has adjusted for measurement error.

In this example, however, because the rate of genotyping error is small (less than 1 per cent), there is little differential effect of measurement error between the two genotypes. That is, if we take the ratio of the slopes as a measure of interaction, then the ratio changes only from 3.33 before correction to 3.54 after correction. If, however, the genotyping error rate had been 3 per cent and the same data had been observed, then the adjusted slopes in the two groups would have been -0.86 and -3.65, respectively, giving a ratio of 4.25. A genotyping error rate of 3 per cent is well within the range often quoted for SNP typing [15]. Moreover, the standard error of the ratio increases with increments in the genotyping error rate. Thus, the combination of errors in both genotyping and exposure assessment can lead to substantial distortion of the interaction effect. Since the bias due to measurement and genotyping errors can be rectified by means of adequate validation studies, future studies could focus on the effect of measurement error on the precision with which interaction parameters are estimated. As can be seen in Table II, poor measurement leads to severe loss of information. Reducing measurement error may often be a preferable strategy than simply increasing the sample size in order to estimate accurately the heterogeneity of risk in a population.

In the future, interventions may increasingly be focused on high risk groups in the population on the basis of studies assessing gene–environment interactions. To be soundly based, such intervention strategies clearly will require unbiased estimation of the combined effects of genetic and environmental factors. Ignoring the effect of even a rather slight error in genotyping on the estimation of the interaction when the exposure is measured with substantial error can be appreciable.

ACKNOWLEDGEMENT

The first author was funded by the Royal Society to conduct this collaborative work. NJW is an MRC Clinician Scientist Fellow.

REFERENCES

1. Luan JA, Wong MY, Day NE, Wareham NJ. Sample size determination for studies of gene–environment interaction. *International Journal of Epidemiology* 2001; **30**:1035–1040.
2. Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene–environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology*, 2003; **32**:51–57.
3. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*. Oxford University Press: Oxford, 1993.
4. Wong MY, Day NE, Bashir SA, Duffy SW. Measurement error in epidemiology: the design of validation studies I: univariate situation. *Statistics in Medicine* 1999; **18**:2815–2829.
5. Wong MY, Day NE, Wareham NJ. Measurement error in epidemiology: the design of validation studies II: bivariate situation. *Statistics in Medicine* 1999; **18**:2830–2845.
6. Duffy SW, Rohan TE, Day NE. Misclassification in more than one factor in a case-control study: a combination of mantel-haenszel and maximum likelihood approaches. *Statistics in Medicine* 1989; **8**:1529–1536.
7. Duffy SW, Maximovitch DM, Day NE. External validation, repeat determination, and precision of risk estimation in misclassified exposure data in epidemiology. *Journal of Epidemiology and Community Health* 1992; **46**: 620–624.
8. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* 1989; **8**:101–169.

9. Qizibash N, Duffy SW, Rohan TE. Repeat measurement of case-control data: correcting risk estimates for misclassification due to regression dilution of lipids in transient ischaemic attacks and minor ischaemic strokes. *American Journal of Epidemiology* 1991; **133**:832–838
10. Gordon D, Ott J. Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pacific Symposium on Biocomputing* 2001; **6**:18–29.
11. Mitchell AA, Cutler DJ, Chakravarti A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics* 2002; **72**: 598–610.
12. Sobel E, Papp JC, Lange K. Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics* 2002; **70**:496–508.
13. Williams DR, Wareham NJ, Brown DC, Byrne CD, Clark PMS, Cox BD, Cox LJ, Day NE, Hales CN, Palmer CR, Shackleton JR, Wang TWM. Undiagnosed glucose intolerance in the community: the Isle of Ely Diabetes Project. *Diabetic Medicine* 1995; **12**:30–35.
14. Wareham NJ, Byrne CD, Williams R, Day NE, Hales CN. Fasting proinsulin concentrations predict the development of type 2 diabetes. *Diabetes Care* 1999; **22**:262–270.
15. Wareham NJ, Wong MY, Hennings S, Mitchell J, Rennie K, Cruickshank K, Day NE. Quantifying the association between habitual energy expenditure and blood pressure. *International Journal of Epidemiology* 2000; **29**:655–660.
16. Wareham NJ, Hennings SJ, Prentice AM, Day NE. Feasibility of heart-rate monitoring to estimate total level and pattern of energy expenditure: the Ely young cohort feasibility study 1994-5. *British Journal of Nutrition* 1997; **78**:889–900.

