Gilbert, R., Martin, R. M., Donovan, J., Lane, J. A., Hamdy, F., Neal, D. E., & Metcalfe, C. (2016). Misclassification of outcome in case-control studies: Methods for sensitivity analysis. *Statistical Methods in Medical Research*, *25*(5), 2377-2393. https://doi.org/10.1177/0962280214523192

MISCLASSIFICATION OF OUTCOME IN CASE-CONTROL STUDIES: METHODS FOR SENSITIVITY ANALYSIS

Rebecca Gilbert[1], Richard M Martin[1], Jenny Donovan[1], J Athene Lane[1], Freddie Hamdy[2], David E Neal[3], Chris Metcalfe[1]

[1] School of Social and Community Medicine, University of Bristol, Bristol, UK

[2] Nuffield Department of Surgery, University of Oxford, Oxford, UK

[3] Department of Oncology, University of Cambridge, Cambridge, UK

Corresponding author: Rebecca Gilbert, School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK. Email: Becky.Gilbert@bristol.ac.uk

Keywords:
Misclassification of outcome; case-control study; risk factors for prostate cancer

Abstract

Case-control studies are potentially open to misclassification of disease outcome which may be unrelated to risk factor exposure (non-differential), thus underestimating associations, or related to risk factor exposure (differential), thus causing more serious bias.

We conducted a systematic literature review for methods of adjusting for outcome misclassification in case-control studies. We also applied methods to simulated data with known outcome misclassification to assess performance of these methods. Finally, real data from the Prostate Testing for Cancer and Treatment (ProtecT) randomised controlled trial, gauged the usefulness of these methods.

Adjustment methods range from recalculating cell frequencies to probabilistic sensitivity modelling and Bayesian models, which incorporate uncertainty in sensitivity and specificity estimates. Simulated data indicated that substantial bias in either direction resulted from differential misclassification. More sophisticated methods, incorporating uncertainty into estimates of misclassification, provided appropriately wide confidence intervals for corrected estimates of risk factor–disease association.

Method choice depends on whether the objective is to assess if an observed association can be explained by bias, or to provide a “corrected” estimate for the primary analysis. Accurate estimation of the degree of misclassification is important for the latter; otherwise further bias may be introduced.

1 INTRODUCTION

Potential risk factors for many important diseases are widely researched using case-control study designs. Associations are investigated by determining whether levels of exposure to specific risk factors differ in cases versus controls.[1] Case-control status may be misclassified, e.g. when a definitive but invasive diagnostic test is carried out only on individuals shown to be at high risk of having the disease by a screening test. For example, in case-control studies that investigate risk factors for prostate cancer (the motivating example used in this paper), there are two main scenarios that may lead to misclassification of outcome:

1) A man with prostate cancer has a prostate specific antigen (PSA) level (a preliminary diagnostic test) below the threshold for further diagnostic evaluation so no biopsy (the definitive test) is carried out, i.e. some men with a 'normal' PSA have prostate cancer.[2]
2) A man may have a negative biopsy when prostate cancer is in fact present. Biopsies sample varying amounts of tissue from different areas of the prostate; the greater the number of cores sampled, the greater the likelihood of finding cancer.[3]

In both scenarios, a 'true' case would be incorrectly classified as a control, i.e. a false negative. In Thompson, 2004[2] 15% of men with PSA<4ng/mL (the usual threshold for biopsy) had prostate cancer. False positives are very unlikely to occur in this context since biopsies will not be positive if no cancer is present.[3] Other contexts may present scenarios where false positives are much more likely, for example, research into predictors of smoking cessation. It may be a concern that, among participants that did not quit, some might report that they had (i.e. specificity<100%).

Misclassification of disease outcome may be unrelated to risk factor exposure (non-differential), when associations will be modestly underestimated,[4-7] or related to risk factor exposure (differential), causing more serious under- or over-estimation of association.[5, 8-10] Misclassification bias in binary outcomes can be described as a function of two parameters, sensitivity and specificity, which can be allowed to differ by exposure group to account for differential misclassification. Sensitivity is the proportion of true positives correctly identified as such and specificity is the proportion of true negatives correctly identified as such.[1] The effect of misclassification of exposure status has been well researched,[8, 11-13] but there has been less methodological work looking at the effects of misclassification of disease status.[6, 7]

If sensitivity and specificity can be accurately estimated, then adjusted estimates of the risk factor-disease association can be presented alongside observed estimates. If assumed ranges of sensitivity and specificity can be justified then simulations can be used to investigate the effect of misclassification on estimates by varying the levels of sensitivity and specificity. Deciding between the two options will depend on the available information and the aim of the analysis. If the researcher is confident in their estimates of sensitivity and specificity, the adjusted estimate could be presented alongside the observed estimate. Varying the amount of sensitivity and specificity may be more appropriate to investigate how the results would change under differing amounts of misclassification and to allow confidence in the direction of the effect.

This paper aims to review and compare the available methods for correcting for non-differential or differential misclassification of outcome in case-control studies. The search for risk factors for prostate cancer will be used as an example area of application. The paper is organised as follows: Section 2 describes the systematic review of the methods available for correcting such misclassification, with results from a simulation study presented in section 3. Section 4 presents the application of the methods demonstrated using the ProtecT study dataset, and in section 5 the findings are discussed and recommendations are made for practical use of these methods.

2 SYSTEMATIC LITERATURE REVIEW

2.1 Review Methods

A systematic literature review was carried out using a MEDLINE search up to 21st January 2010 to identify papers that present methods for adjusting for misclassification of outcome in case-control studies. The following search strategy was used, where one term from each group must be present:

- Sensitivity and Specificity (MeSH); disease misclassification, sensitivity, specificity, misclassification, misclassification bias, differential misclassification, nondifferential misclassification, non differential misclassification (text)
- Logistic Models (MeSH); logistic regression, binomial regression, case-control studies, binary outcome, odds ratio, Odds Ratio (text)
- (Prostatic Neoplasms (MeSH); prostate cancer, outcome, disease (text).

Extra search terms regarding prostate cancer were included in the search, as this is the area of application of interest to us. Results were limited to the English language. An automatic alert was set up for the same search and any new studies were incorporated up until 23rd March 2012. Papers were

included in the review if they presented methods for dealing with outcome misclassification in case-control studies of any outcome. Papers were excluded from the review if they considered exposure or covariate misclassification, other types of bias such as recall or selection bias, presented methods for adjusting rate ratios, or presented methods for adjusting odds ratios from cohort studies. Papers that presented supporting information, such as information regarding misclassification of prostate cancer status, were also retrieved and referenced as appropriate.

The initial search identified 9 199 hits, of which 8 893 were in English. After reading titles and abstracts for potential relevance, 36 papers were retrieved for detailed reading. Sixteen papers were included, of which seven were methodological.[4-7, 14-16] Searching the reference lists of these sixteen papers led to a further twelve papers being retrieved and seven papers being included, of which four were methods.[8, 17-19] A textbook that presented a method was also identified via a reference list and included as a method.[20] Citation searches, using ISI Web of Knowledge, were carried out on six 'key' papers,[4-7, 15, 19] leading to a further eighteen papers being retrieved and a further five papers being included, one being a method.[21] Personal communications led to one further paper being added, which has since been published.[22] The automated alert (up to 27th May 2013) identified one further paper for inclusion.[23] Overall, fifteen methodology papers were included in the review (**Figure 1**). Web of Science was also searched with a similar set of terms, identifying 5155 titles and abstracts. Screening these did not identify any new papers.

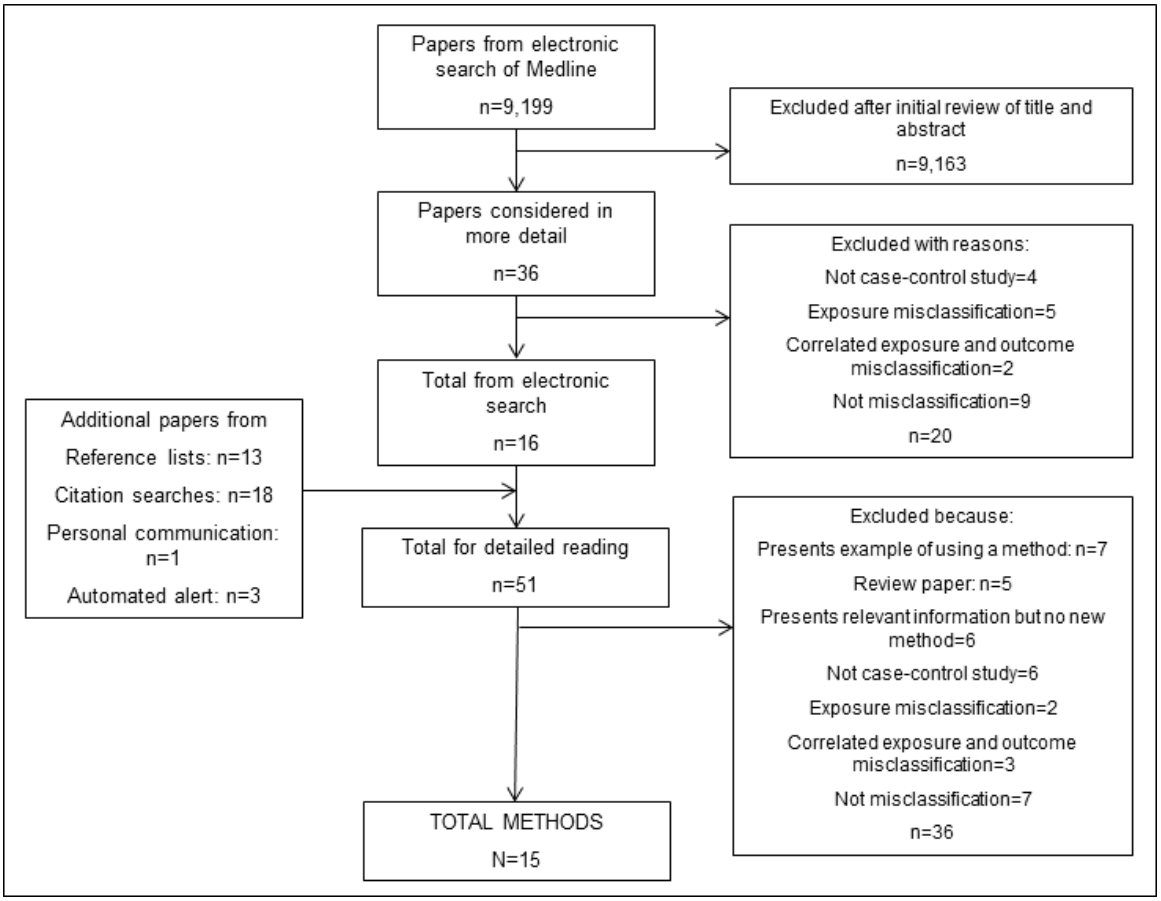

Figure 1: Flowchart of included papers (up to 27th May 2013)

2.2 Notation

Methods fall into five main categories, with the following notation used to describe each (**Box 1**).

The observed (biased) numbers and odds ratio shall be denoted by the superscript *.

The standard 2*2 table will be denoted as:

	Cases (D_1)	Controls (D_0)	
Exposed (E_1)	a	c	M_1
Not exposed (E_0)	b	d	M_0
	A	B	N=A+B

$$\Rightarrow \mathrm{OR} = ad/bc$$

Sensitivity and specificity shall be denoted as Se_i and Sp_i respectively, where i = 1 for the exposed group and i=0 for the unexposed group and can be calculated in the following way:

Classified disease status	Actual Disease Status D_{1i}	D_{0i}	
D_{1i}*	r_i	t_i	
D_{0i}*	s_i	u_i	
	R_i	T_i	N_i

Se_i = Sensitivity in i = $r_i/(r_i + s_i)$;

Sp_i = Specificity in i= $u_i/(t_i + u_i)$

Where:

Under non-differential misclassification $Se_1=Se_0$ and $Sp_1=Sp_0$.

$E[Se_i + Sp_i - 1] \geq 0$

VSe_i and VSp_i are the variances of Se_i and Sp_i respectively, where

$VSe_i = Se_i*(1- Se_i)/R_i$ and $VSp_i = Sp_i*(1- Sp_i)/T_i$;

$P(D_{0i})$ and $P(D_{1i})$ are the actual proportions of subjects in exposure group i who are controls and cases respectively;

$P(D^*_{0i})$ and $P(D^*_{1i})$ are the proportion of subjects in exposure group i observed as controls and cases respectively.

Box 1: Notation

2.3 Altering cell frequencies to adjust the crude estimate of association

Eight papers identified present alternative methods which alter the cell frequencies of a 2*2 table. Four papers demonstrate the effect of outcome misclassification by estimating the percentage of misclassification and altering the cell frequencies accordingly.[5-7, 22] Another four methods use estimates of sensitivity and specificity to calculate the true numbers of diseased and non-diseased.[14, 17, 18, 20] The new cell counts can then be used to calculate odds ratios as usual. These methods can be easily applied to published summary statistics. Differential misclassification can be incorporated by allowing the sensitivity and specificity to differ by exposure group.

For example, Lash et al[20] present the equations in **Table 1** for recalculating the cell frequencies based on estimates of sensitivity and specificity. If $Sp_1 = Sp_0 = Se_1 = Se_0 = 1$, the adjusted cell frequencies are equal to the observed cell frequencies.

	Observed		Adjusted	
	D_1*	D_0*	D_1^I	D_0^I
E_1	a	c	$a^I = [a-(a+c)(1-Sp_1)]/[Se_1 - (1-Sp_1)]$	$c^I = (a+c) - a^I$
E_0	b	d	$b^I = [b-(b+d)(1-Sp_0)]/[Se_0 - (1-Sp_0)]$	$d^I = (b+d) - b^I$
Total	A=a+b	B=c+d	$A^I = a^I + b^I$	$B^I = c^I + d^I$

Table 1: Formulae for correcting observed data for estimated values of sensitivity and specificity (based on Lash, 2009[20])

2.3.1 Adjusted Standard Errors

Sensitivity and specificity can be estimated either 'internally' from a validation sub-sample or from comparable external data. When recalculating odds ratios based on recalculated cell frequencies, it is not appropriate to calculate confidence intervals in the standard way if sensitivity and specificity have been estimated from external data. Standard confidence intervals only take account of sampling variation and would be too narrow. Methods for incorporating systematic error into the calculation of standard error, when external estimates of sensitivity and specificity have been used to adjust for misclassification of outcome, are presented by Greenland et al, 1988.[19] The variance of the log odds ratio, incorporating estimates of sensitivity and specificity, can be estimated using the following equations:

$$Var(non-differential) = \frac{VSe({}^{1}/_{P(D_{01})} - {}^{1}/_{P(D_{00})})^2 + VSp({}^{1}/_{P(D_{11})} - {}^{1}/_{P(D_{10})})^2 + \sum_{i=0}^{1} P(D_{0i}^*)P(D_{1i}^*)/M_i P(D_{0i})^2 P(D_{1i})^2}{(Se + Sp - 1)^2} \quad (1)$$

$$Var(differential) = \sum_{i=0}^{1} \frac{VSe_i/P(D_{0i})^2 + VSp_i/P(D_{1i})^2 + P(D_{0i}^*)P(D_{1i}^*)/M_i P(D_{0i})^2 P(D_{1i})^2}{(Se_i + Sp_i - 1)^2} \quad (2)$$

The equations for the two scenarios differ since, under non-differential misclassification, the corrected case numbers are not independent of the corrected control numbers. In the above equations, $P(D_{0i})$ and $P(D_{1i})$ (the actual proportions of subjects in exposure group i who are controls and cases

respectively) can be calculated from one of the presented methods for recalculating cell frequencies, for example Lash et al[20] as discussed in section 2.3.

2.4 Logistic Regression Models

So far, all of the considered methods have made corrections at the summary statistic/aggregate data level. A method that can be applied to individual records enables complex relationships involving adjustment for confounders to be investigated. Magder et al[15] presents a method that fits a logistic regression model with each subject being included twice, as diseased and not diseased, with weights determined by the probability that the subject is truly diseased given the data. The procedure is described using the following example.

Suppose an individual, classified as diseased, has 90% probability of being truly diseased, determined by sensitivity, specificity and values of the individual's covariates reflecting a known risk factor. This individual would be entered into standard logistic regression twice, once as diseased with a weight of 0.9 and once as not diseased with a weight of 0.1. Since these probabilities depend partly on the values of covariates and therefore on the parameters of the logistic regression, they need to be recalculated after the logistic regression parameters are estimated. This process is repeated using the new probabilities as weights until the model converges. Differential misclassification can be accommodated by allowing the values of sensitivity and specificity to differ for individuals. The formal algorithm is presented in the paper, a downloadable macro has been implemented in the SAS software package, and a user-written command ('logitem') is available for the Stata software package/

This method cannot be applied to simple case-control studies since there is no way to estimate the underlying probability of disease as the numbers of cases and controls are fixed by design. One way around this would be to incorporate external estimates of disease frequency as sampling weights.

2.5 Probabilistic Sensitivity Analysis

Probabilistic sensitivity analyses is a semi-Bayesian approach, a compromise between classical and Bayesian methods which allow prior distributions to be assigned to the unknown parameters, i.e. sensitivity and specificity estimates. This produces a frequency distribution of adjusted estimates of the association which can be easily summarised by a median and confidence intervals that

incorporate both random and systematic error. This differs from traditional Bayesian methods, where distributions are assigned to all parameters and specialist software is required to refit models.

For example, the method discussed earlier by Lash et al[20] (**Table 1**) can be incorporated into a probabilistic sensitivity analysis. The user can define probability distributions of sensitivity and specificity (either two distributions for non-differential misclassification or four distributions for differential misclassification). Since the sensitivity and specificity in the exposed group is likely to be related to the sensitivity and specificity in the non-exposed group if the same diagnostic procedure has been used in each, a correlation coefficient between the sensitivity/specificity in the non-exposed group and the exposed group must be specified (a correlation of 1 indicates non-differential misclassification by forcing the sensitivity/specificity in the two groups to be identical). The method assumes that the true misclassification is non-differential, which may be observed by chance as differential misclassification (due to the specified correlation between the sensitivities/specificities in the two groups). Therefore, the method cannot correct for true differential misclassification.

Estimates of sensitivity and specificity, by exposure group, are randomly sampled from the defined distributions.[20] The formula given in **Table 1** can then be applied substituting in the random values of sensitivity and specificity calculated above, and an adjusted odds ratio is calculated (OR^{I}), accounting for systematic error. A further step calculates an odds ratio corrected for both systematic and random error (OR^{II}). The additional random error is calculated from multiplying the observed standard error by a random number, r, r~N(0,1):

$$OR^{II} = \exp\left[\ln(OR^{I}) - \sqrt{Var} * r\right]$$

This process is repeated many times to create a distribution of odds ratios, corrected for misclassification. Impossible values are dropped (for example, combinations of sensitivity and specificity that lead to negative cell counts). Results are easily summarised as a median estimate of association and confidence intervals that can be compared to the observed results. Three intervals can be outputted: the conventional 95% confidence interval (accounting for random error), a simulation interval that accounts for systematic error, and a second simulation interval that accounts for both random and systematic (total) error. Macros to carry out this method are available for SAS

and Excel via the author's website (http://sites.google.com/site/biasanalysis, accessed 18/11/2011). Alternative methods are available which correct the individual data values (Fox et al,[8] Lyles et al[23]).

2.6 Bayesian Methods

A fully Bayesian method (McInturff et al[16]) uses subjective prior information, for example, estimates of sensitivity and specificity, to estimate corrected log odds ratios under a logistic regression model when there is non-differential misclassification. Sensitivity and specificity are not treated as fixed values, and uncertainty is allowed in the estimates of them. A simple Bayesian method for logistic regression is applied using the WinBUGS software. The likelihood function is written in terms of the coefficients, sensitivity and specificity:

$$L(\beta, Se, Sp) = \prod_{j=1}^{n} [\pi_j Se + (1-\pi_j)(1-Sp)]^{y_j} [\pi_j (1-Se) + (1-\pi_j) Sp]^{1-y_j} \quad (3)$$

where β is a vector of regression coefficients; y_j denotes the observed outcome for the jth individual, j=1,…,n; and $\pi_j = \Pr(z_j = 1|x_j)$ is the probability that an individual is diseased, based on the observed outcomes across individuals with the same risk factor profile.

Independent beta priors, Be(a,b), are assumed for sensitivity and specificity, where a-1 successes are expected out of a+b-2 trials. A multivariate normal distribution is assumed for the coefficients. The joint posterior is:

$$p(\beta, Se, Sp|X, Y) \propto L(\beta, Se, Sp) p(\beta) p(Se) p(Sp)$$

and Gibbs sampling is used to obtain a numerical approximation to the posterior distributions, from which adjusted odds ratios can be calculated. The WinBUGS code is given in the appendix to the paper.[16] This method assumes that covariates are measured perfectly, and cannot accommodate differential misclassification. It is not included in the Simulation Study presented here, which focuses on differential misclassification, but is applied to the real data example in section 4. An alternative Bayesian approach is presented by Gerlach et al.[21]

3 SIMULATION STUDY

A simulation study was carried out, firstly, to assess how odds ratios are affected by misclassification and, secondly, to assess how well the presented methods adjust effect estimates for misclassification. The data are designed to reflect an exposure that is inversely associated with prostate cancer

(specifically, diabetes, as presented in section 4) which is expected to be related to (differential) misclassification of outcome. The mean odds ratio prior to misclassification being incorporated (i.e. the unbiased estimates) is OR=0.79, 95% CI: 0.64, 0.97 (**Box**).

Two-hundred simulated datasets were created, based on the distribution of diabetes exposure in the real data example in section 4. All methods were applied to the same 200 datasets. The number of true cases and true controls who are exposed to the risk factor (n_{11} and n_{01} respectively) is generated using a binomial distribution with parameters ~Bin(number, probability of being exposed), where the parameters are taken from the distribution as presented in the real data example. Misclassification is assumed to be due to less than perfect sensitivity and specificity of the diagnostic process, and that there are no other biases present. Hence, 'observed' datasets with a known mean sensitivity were created, i.e. by recoding a proportion of true cases as observed controls. Specificity is assumed to be 100% under all scenarios, in line with the cancer diagnosis example where cancer would not be diagnosed from biopsies in its absence. As presented in Box 2, the number of observed cases is π_{11} amongst the exposed and π_{10} amongst the unexposed. The notation and distributions are shown in **Box 2**. All simulations are based on random numbers generated using the Stata '`rbinomial`' function.

Characteristics of the 'true' simulated data

N=13 175 (1 733 cases)

Within cases, the number exposed is randomly realised by $n_{11} \sim Bin(1733, 0.065)$ and therefore $n_{10} = 1733 - n_{11}$

Within controls, the number exposed is: $n_{01} \sim Bin(11442, 0.081)$ and therefore $n_{00} = 11442 - n_{01}$

$OR = \frac{0.065/(1-0.065)}{0.081/(1-0.081)} = 0.7887$

Notation and distributions for simulated data

	Truth				Observed (biased)		
	D_1	D_0			D_1	D_0	
E_1	n_{11}	n_{01}		E_1	$\pi_{11} \sim Bin(n_{11}, Se_1)$	$\pi_{01} = (n_{11} + n_{01}) - \pi_{11}$	
E_0	n_{10}	n_{00}		E_0	$\pi_{10} \sim Bin(n_{10}, Se_0)$	$\pi_{00} = (n_{10} + n_{00}) - \pi_{10}$	
	n_1	n_0	N		n_1	n_0	N

For non-differential misclassification: $Se_1 = Se_0$

Box 2: Characteristics of the 'true' simulated datasets

Each method from section 2 was evaluated by comparing the mean of the corrected beta coefficients (across simulated datasets) to the true beta value, and by comparing the mean method-specific standard error (calculated for each corrected beta distribution) to the empirical standard deviation (the standard deviation of estimated beta coefficients across simulations).

3.1 Results

3.1.1 Characteristics of the simulation data

The change in mean odds ratio with varying levels of sensitivity showed potential for both the magnitude and direction to be incorrectly estimated depending on the amount of misclassification present (**Figure 2**). Examining the estimated odds ratios when the misclassification is non-differential, there is the expected attenuation towards the null result, but it is very small in magnitude. Under differential misclassification, for a given difference between sensitivity in the exposed and unexposed groups (for example, a difference of 10%), there is a constant under- or over-estimation of the odds ratio.

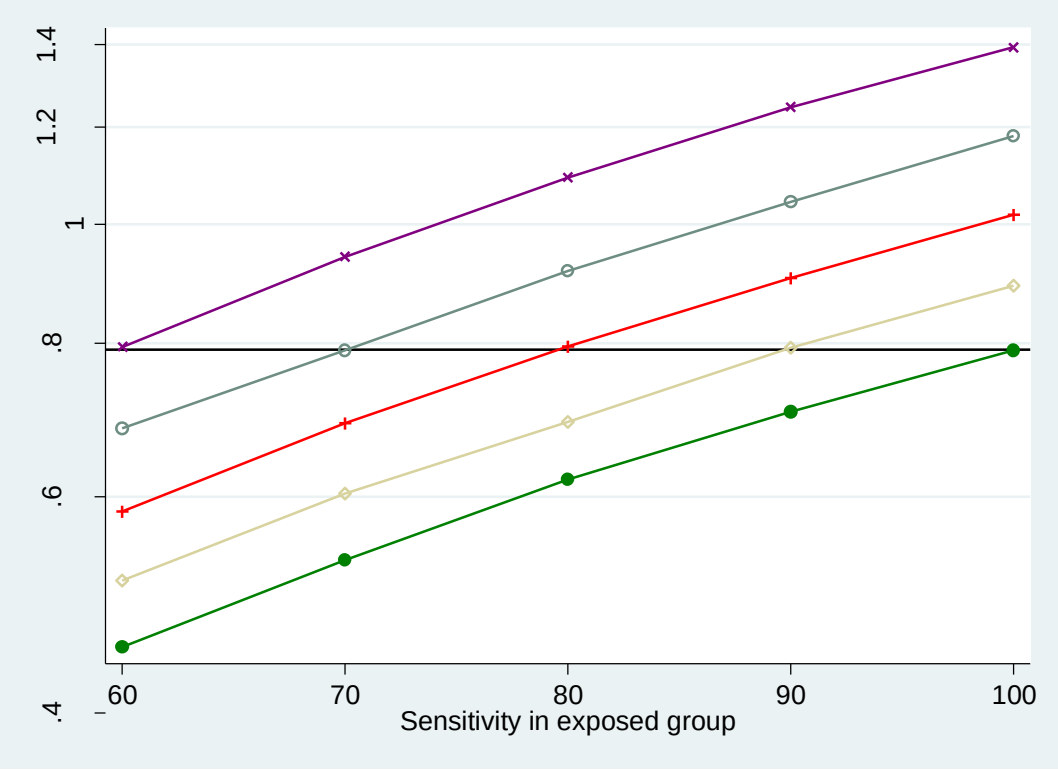

Figure 2: Change in mean odds ratio (presented on a log scale) by varying sensitivity in the exposed and unexposed groups. The odds ratios under varying levels of sensitivity in the unexposed group are

depicted by: x 60%; ○ 70%; + 80%; ◊ 90%l; ● 100%. The true OR is depicted by the line parallel to the x-axis.

3.2 Comparison of methods using simulated data

Table 2 gives the results with the different methods of correcting for misclassification of outcome. Under the right assumptions, the method of altering tabulated frequencies by estimating sensitivity corrects back to the true OR, but as the method is applied using increasingly wrong estimates of sensitivity, the corrected OR begins to differ quite widely from the 'true' OR. Very similar results are obtained using logistic regression methods incorporating sensitivity estimates.

Therefore, these methods can correct results accurately if there are accurate estimates of sensitivity. The simulation study was repeated for a non-differential example (data not shown) and, as expected, providing that the assumption of non-differential misclassification was correct, then no bias was introduced even if the wrong amount of misclassification was assumed.

Table 2: Mean log odds ratios (ln(OR)), mean standard error and empirical standard deviation produced for differential misclassification (with known true sensitivity of 90% in exposed and 80% in non-exposed groups), after adjusting for varying sensitivities using: altering cell frequencies using estimated sensitivity (Lash et al[20]); Logistic regression models (Magder et al[15]); Probabilistic sensitivity analysis using the method from Lash et al[20]) . Mean corrected standard error calculating using method by Greenland et al[19]. Specificity=100%.

Assumed Sensitivity (%)		**Mean ln(OR)**	**Empirical SD**	**Mean Std Error**	**Mean corrected Std Error**
Exposed	**Unexposed**				
True ln(OR)		-0.24		0.08	
Altering cell frequencies using estimated sensitivity					
70	80	0.04	0.12	0.09	0.12
80	80	-0.11	0.12	0.10	0.11
90	80	-0.24	0.11	0.10	0.11
90	90	-0.11	0.11	0.10	0.11
Logistic regression models					
70	80	0.04	0.12	0.12	
80	80	-0.11	0.12	0.10	
90	80	-0.24	0.11	0.09	
90	90	-0.11	0.11	0.10	
Probabilistic sensitivity analysis					
Sensitivity Range (%)	**Corr (Se_1,Se_0)**		**a**	**b**	
60-100	0.8	0.11	0.20	0.19	
50-90	0.8	-0.11	0.21	0.21	
50-90	0.6	-0.11	0.25	0.25	

[a] Estimated as the standard deviation of all the beta coefficients that make up the distribution.
[b] Estimated via calculating the error factor from the pseudo confidence interval.

The frequency distributions of adjusted odds ratios, for three different 'observed' scenarios corrected for estimates of sensitivity, using probabilistic sensitivity analysis to assign probability distributions to estimates of sensitivity are presented in **Table 2**. Within scenarios (the columns), the estimate varies more. However, across scenarios (the rows), the mean estimate varies less as assumptions regarding the direction of differential misclassification are not incorporated. This method does not adjust for 'true' differential misclassification, but allows for chance differential misclassification in the observed data.

The mean beta coefficients, empirical standard deviation and mean standard errors collected after applying the different methods to correct for varying estimates of misclassification, assuming that the correct sensitivity is 90% in the exposed group and 80% in the non-exposed group can be seen in **Table 2**. The standard error for data without misclassification is 0.08. After adjusting for misclassification, a larger standard error would be expected, that incorporates the extra error introduced by having to estimate the amount of misclassification. This is indicated by the observed variation in estimates across 200 simulated datasets with random error and misclassification. Under the differential scenario, the standard errors when using estimates of sensitivity to alter table cell frequencies are slightly increased (mean standard error=0.10) due to differences in observed and adjusted proportions. The formula to correct the standard error using estimates of sensitivity and specificity, gives an estimated standard error of 0.11, whilst the logistic regression model gives an estimated standard error of 0.12, both in close agreement with the SD. With vague prior knowledge the probabilistic sensitivity analysis gives estimates which vary more across simulated datasets. The model-based SE's estimate this variation accurately.

4 APPLICATION OF METHODS TO PROTECT DATA

4.1 Example dataset: ProtecT

During recruitment to the ProtecT study (between 2001 and 2009), men aged 50-69 years at 400 general practices in nine UK centres were offered a PSA test at a community-based 'prostate check clinic', and those with raised levels ($\geq$ 3 ng/ml) were offered diagnostic 10-core biopsy.[24] All

participants in ProtecT who had no evidence of prostate cancer were eligible for selection as controls for nested case-control studies of risk factors; that is, men with a PSA test < 3ng/ml or a raised PSA (≥ 3 ng/ml) combined with at least one negative biopsy and no subsequent prostate cancer diagnosis during the follow-up protocol for negative biopsies.

Cancer Research UK includes family history of prostate cancer and diabetes mellitus amongst the risk factors for prostate cancer.[25] Family history can increase the risk between two and seven-fold, depending on the age of onset and the number of relatives affected.[25] A meta-analysis of studies assessing the association between diabetes and prostate cancer found that people with diabetes had a 9-16% decrease in risk of developing prostate cancer.[26]

Family history has been shown not to be associated with PSA level, and so misclassification of outcome will be non-differential. It has been suggested that PSA acts as a distorter variable,[27] in that if the risk factor of interest is associated only with PSA level, it will also appear to be associated with prostate cancer if case-control status is determined in men with a high PSA level. Diabetes is associated with lower PSA levels,[28, 29] so diabetic men with prostate cancer may be less likely to undergo biopsy and hence have their prostate cancer detected. This potentially creates an artificial inverse association between diabetes and prostate cancer. There is therefore potential for differential misclassification between men with and without diabetes, as, in men with diabetes, cases are more likely to be misclassified as controls.

4.2 Estimates of sensitivity and specificity

Recent estimates obtained from the placebo arm of the US Prostate Cancer Prevention Trial (PCPT)[30] estimated a sensitivity of 32.2% and specificity of 86.7% using a PSA threshold of 3.1ng/mL. Due to repeated screening in the US, it is likely that sensitivity is under-estimated, since cases are more likely to be diagnosed at an early stage of disease progression (i.e. cases with high PSA and large tumours will have been removed from the cohort). Sensitivity did increase in men judged to have clinically important disease: using a PSA cut=off of 3.1ng/mL gave a sensitivity of up to 68.4% for men with tumours judged to be aggressive when compared with less aggressive or no cancer.

The true sensitivity of biopsy (alone, without considering PSA testing) is also difficult to determine, since men with low PSA levels do not often undergo biopsy, and men who have a negative biopsy

result do not undergo surgical verification. Estimates of the sensitivity of biopsy range from 30-80%, depending on the number of cores and the zones biopsied.[3, 31, 32] The specificity of biopsy is consistently approximately 100%, i.e. it is very rare to mistake other conditions for cancer.[3, 31, 32]

A plausible range of sensitivities under which to investigate how the estimates may be affected if misclassification is present in the current example is assumed to be 60-100%. Diabetes may be associated with increased obesity and prostate volume, making biopsies more difficult and therefore lowering sensitivity in men with diabetes compared to men without diabetes. For ease of presenting results, specificity is held equal to 100%.

4.3 Results

The observed results are presented in **Table 3**. Recall bias is unlikely, since details of diabetes and family history were collected from men prior to their PSA results or subsequent biopsy results becoming available.

Family History of Prostate Cancer	Cases	Controls	Diabetes	Cases	Controls
Yes	205	930	Yes	113	923
No	2 167	16 086	No	1 620	10 519
Total	2 372	17 016	Total	1 733	11 442
	$\Rightarrow OR^* = 1.64$			$\Rightarrow OR^* = 0.79$	
	95% CI: 1.40, 1.92			95% CI: 0.65, 0.97	
	p-value<0.001			p-value=0.026	

Table 3: Observed Results from the ProtecT Data

4.3.1 Altering cell frequencies and logistic regression model

The sensitivity in the exposed and non-exposed groups was allowed to vary between 60-100% (whilst holding specificity equal to 100%) using the method presented by Lash et al[20] (section 2.3). The odds

ratios for family history slightly increase with increasing non-differential misclassification from the observed result of 1.64 to a corrected result of 1.75 when corrected for underestimation due to non-differential misclassification, with sensitivity equal to 60% (**Table 4**). For diabetes (**Figure 4**), the odds ratio decrease remains constant under increasing non-differential misclassification from the observed result of 0.79 to 0.78 when sensitivity is 60%. Under assumed differential misclassification, however, the odds ratios vary from 0.43 (Se_1=100, Se_0=60) to 1.44 (Se_1=60, Se_0=100). With sensitivity estimated as 60% for men with and 65% for men without diabetes, the odds ratio is corrected to 0.86 (0.86,1.09; Table 4). The results using the logistic regression method showed an almost identical pattern (**Table 4**).

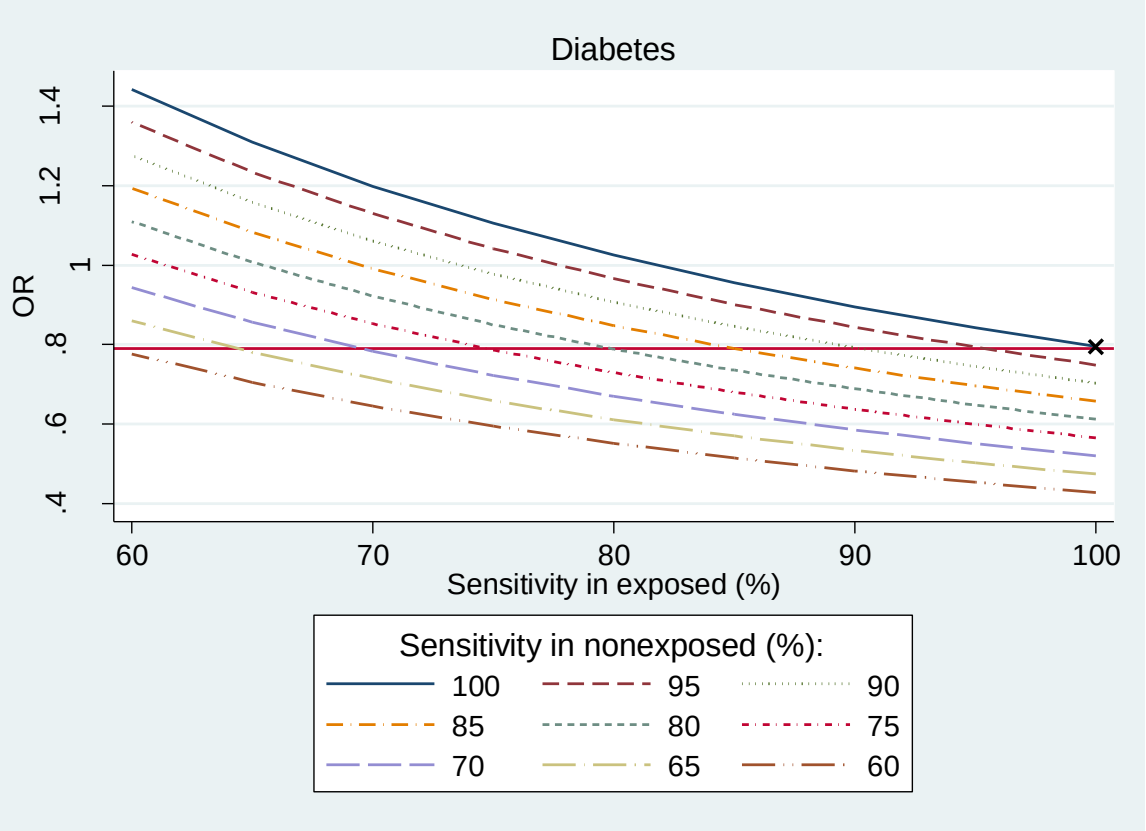

Figure 4: Correcting for misclassification using estimates of sensitivity and specificity using Lash, 2009[33] , with specificity held at 100%. X and the horizontal line show the observed result.

4.3.2 Probabilistic Sensitivity Analyses

The probability sensitivity method was applied to the simple method presented by Lash et al[20] (**Table 1**), allowing both Se1 and Se0 to range between 60 and 100%, whilst holding Sp1 and Sp0 equal to 100%, by randomly sampling from a uniform distribution, [0,1), and allowing 100 000 iterations. For differential misclassification, the correlation coefficient for sensitivities and specificities was equal to ρ=0.8. For family history, the median estimate and 95% confidence intervals incorporating only systematic error are OR=1.67 (1.64,1.74) (**Table 4**). After incorporating total error, the results are OR=1.68 (1.43,1.97). For diabetes, incorporating systematic error gives a median OR of 0.79 (0.61,1.01). Incorporating total error gives a median OR of 0.79 (0.57,1.08).

4.3.3 Bayesian Methods

The Bayesian model provided by McInturff[16] was applied to the data using a prior distribution for sensitivity of beta(a,b), where there are (a-1) successes out of (a+b-2) trials with an initial value of a=80%, i.e. beta(81,21). Specificity was fixed at 100%. Vague prior distributions were assigned to the regression coefficients (normal(0,0.001)), and the initial values, based on the beta coefficients from logistic regression models, were intercept β_1=-2 and association β_2=0.5 for family history and β_1=-2; β_2=-0.2 for diabetes. Convergence was ensured by allowing 100,000 burn-in iterations. The following 50 000 iterations were monitored to calculate mean estimates and 95% posterior probability intervals. For family history, the mean estimate and 95% posterior probability intervals are OR=1.67 (1.41,1.98) and for diabetes are OR=0.79 (0.64,0.96) (**Table 4**). The method does not accommodate differential misclassification.

Comparing the different effect estimates for family history (**Table 4**), the overall conclusion would not alter and the association between family history and prostate cancer, under these assumptions, appears to be slightly stronger than the observed results indicate. Considering diabetes, the direction of effect remains the same although the confidence intervals have widened to include the null result. The association between diabetes and prostate cancer, under these assumptions, therefore appears to be slightly weaker than indicated by the observed data, although it is unlikely that the overall conclusion would change in light of these findings.

Table 4: Adjusted effect estimates for two risk factors for prostate cancer when adjusted for misclassification of outcome

	Family History (assuming misclassification is non-differential) **OR (95% CI)**	**Diabetes** (assuming misclassification is differential) **OR (95% CI)**
Observed	1.64 (1.40, 1.92)	0.79 (0.65, 0.97)
Altering cell frequencies using estimates of sensitivity[a]	1.75 (1.45,2.10)[c]	0.86 (0.68,1.09) [c]
Logistic Regression[a]	1.75 (1.45,2.10)	0.86 (0.69,1.07)
Probabilistic Sensitivity Analyses[a,b]	1.75 (1.49,2.06)	0.78 (0.61,1.00)
Bayesian Methods[a]	1.67 (1.41,1.98)	Not applicable

[a] Non-differential sensitivity=60%. Differential sensitivity is 60% for men with diabetes and 65% for men without diabetes
[b] Correlation coefficients for non-differential and differential misclassification are ρ=1 and ρ=0.8 respectively. Sensitivities vary between 50-70% for family history and 55-75% for diabetes, drawn from a uniform distribution.
[c] Adjusted CI, calculated using standard errors for misclassification.

5 DISCUSSION

5.1 Summary of Findings

This paper firstly reviewed the literature for methods of adjusting for misclassification of outcome in case-control studies, which identified sixteen methods. Secondly, simulation datasets were used to assess the performance of these methods. The methods were then applied to data from the ProtecT study by adjusting the effect-estimates of two risk factors for prostate cancer using estimated levels of outcome misclassification.

Empirically adjusting for misclassification by estimating sensitivity produced corrected odds ratios that were very similar to the true odds ratio under both non-differential and differential misclassification when the estimates of sensitivity were accurate. However, misleading results could be produced under differential misclassification if sensitivity is estimated inaccurately, with both magnitude and direction of association being incorrect, although this only occurred under extremely mis-estimated sensitivities.

Standard errors incorporating the extra error can be calculated and used to produce confidence intervals that reflect the extra uncertainty.

The results from the logistic regression model were identical to those from the method of empirically incorporating estimates of sensitivity, despite the stated limitation that the method is not supposed to be suitable for use with simple case-control studies where the number of cases and controls are fixed by design. The estimated standard error was increased to reflect the extra error in estimating sensitivity. Unlike the simpler method, adjustment for confounders could be incorporated into this method as with standard logistic regression.

Using probabilistic sensitivity analysis produced adjusted odds ratios that were not as accurately corrected back to the true odds ratios as with previous methods, but neither did they vary as much under inaccurate estimates of sensitivity. Allowing for a range of sensitivities and specificities allows calculation of simulation intervals that reflect the uncertainty in the starting values. There is therefore the added benefit of confidence intervals that incorporate both random and systematic error.

In our real data example, under likely levels of non-differential misclassification, the conclusion regarding the direction of the association between family history and prostate cancer would not change, although the magnitude does differ slightly. Under the assumption of differential misclassification, the conclusion regarding the associations between diabetes and prostate cancer would not change, although the magnitude changes and some methods cannot rule out a null result or a positive association under some more extreme combinations of sensitivity.

5.2 Applying the Methods

The choice of method depends on the available estimates of sensitivity and specificity, the form of the data (individual data or summary level), and whether the objective is to assess whether an observed association can be explained by bias or to provide an adjusted estimate. If sensitivity and specificity can be accurately estimated, then adjusted estimates of the risk factor-disease association can be presented alongside observed estimates. Varying the amount of sensitivity and specificity may be more appropriate to investigate how the results would change under differing amounts of misclassification. This may provide reassurance, for example, that, even if there is misclassification, the overall conclusion would not change. Methods exist for application at both the record-level or at

the summary data level. Summary data level methods can be used on one's own data or to adjust published results. Record-level techniques may be preferable, since they enable the investigator to adjust for confounders and inter-relations between variables. It is important to provide the ranges, distributions and rationale for all assumptions so that readers can judge themselves whether the methods have been used appropriately.

Estimates of sensitivity and specificity can be gathered from either a validation sample or estimated from other studies with comparable data. However, the degree of misclassification is likely to be study specific, and external estimates taken from other studies may not apply to the current study.[18] It is therefore important to carefully consider where estimates of sensitivity and specificity are derived from. Methods can also be altered to accommodate estimates of the positive predictive value and negative predictive value if estimates of sensitivity and specificity are not available. Estimates of sensitivity and specificity are more likely to be available non-differentially than broken down by exposure group.

A number of authors[34, 35] warn against the overuse of adjusting for misclassification. There should be evidence that misclassification is actually present: mentioning the possibility of a limitation may be used to discount findings without any proof that the limitation is actually present. The assumption of non-differential misclassification should be deduced logically: simulations have demonstrated that adjusting under the assumption of non-differential misclassification introduces extra bias if the misclassification is actually differential[10] (section 3). It is therefore important to make logical assumptions about the type and extent of misclassification before deciding whether to adjust, and what method to use.

Ideally, the study design and classification method would be improved so that no correction is necessary, although this is impractical in most studies. Alternatively, avoiding differential misclassification in favour of non-differential misclassification would allow estimates to be left unadjusted, as fairly large amount of non-differential misclassification cause little problem. Through straightforward algebra it can be shown that non-differential misclassification is equivalent to adding a positive constant to the numerator and denominator of the odds ratio, which will shrink the ratio towards one. Examining the controls to detect any missed disease is a possibility, but not feasible when the diagnostic test, e.g. biopsy, is invasive, unpleasant and associated with side-effects. Other

potential methods for avoiding or reducing misclassification of outcome pose a series of problems.[6] For example, limiting the control group to subjects who have undergone a diagnostic test and been pathologically confirmed to be disease-free would increase detection, thus improve the sensitivity, but also increase the number of unnecessary diagnostic tests. The selection of a pathologically confirmed disease-free control group is likely to reduce the number of eligible subjects, thus reducing the sample size and therefore precision.

5.3 Conclusions

This paper has demonstrated that misclassification of outcome in case-control studies can bias estimates and may lead to incorrect conclusions being drawn. It is therefore important that epidemiological studies of disease-risk factor associations attempt to assess the possible impact of outcome misclassification in their data, and aim to avoid differential misclassification if possible. This paper has presented a number of methods by which this may be carried out and the performance of each described. Since estimates of sensitivity and specificity are required and are often difficult to accurately ascertain, using these methods as useful tools for sensitivity analysis assessing the impact of outcome misclassification on effect estimates over a range of plausible values may be the best approach.

Acknowledgements

This work was funded by a Cancer Research UK Graduate Training Fellowship [RG: C31211/A10095]. The UK Department of Health funded the ProtecT study through the NIHR Health Technology Assessment programme. Trial registration number: ISRCTN20141297. All men provided written informed consent prior to inclusion in the study. Trent Multicentre Research Ethics Committee (MREC) approved the ProtecT study (MREC/01/4/025) and the associated ProMPT study which collected biological material (MREC/01/4/061). Department of Health disclaimer: The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Department of Health. The authors would like to acknowledge the tremendous contribution of all members of the ProtecT study participant and researchers.

Conflict of interest statement: None declared.

Reference List

1. Kirkwood BR and Sterne JAC. *Essential Medical Statistics*. Blackwell Science, 2003.
2. Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *New England Journal of Medicine*. 2004; 350: 2239-46.
3. Haas GP, Delongchamps NB, Jones RF, et al. Needle Biopsies on Autopsy Prostates: Sensitivity of Cancer Detection Based on True Prevalence. *JNCI Journal of the National Cancer Institute*. 2007; 99: 1484-9.
4. Copeland KT, Checkoway H, McMichael AJ and Holbrook RH. Bias due to Misclassification in the Estimation of Relative Risk. *American Journal of Epidemiology*. 1977; 105: 488-95.
5. Godley PA and Schell MJ. Adjusted Odds Ratios Under Nondifferential Misclassification: Application to Prostate Cancer. *Journal of Clinical Epidemiology*. 1999; 52: 129-36.
6. Platz EA, De Marzo AM and Giovannucci E. Prostate Cancer Association Studies: Pitfalls and Solutions to Cancer Misclassification in the PSA Era. *Journal of Cellular Biology*. 2004; 91: 553-71.
7. Chyou PH. Patterns of bias due to differential misclassification by case-control status in a case-control study. *European Journal of Epidemiology*. 2007; 22: 7-17.
8. Fox MP, Lash TL and Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International Journal of Epidemiology*. 2005; 34: 1370-6.
9. Greenland S and Gustafson P. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *American Journal of Epidemiology*. 2006; 164: 63-8.
10. Luan XQ, Pan W, Gerberich SG and Carlin BP. Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Statistics in Medicine*. 2005; 24: 2221-34.
11. Chu H, Wang Z, Cole SR and Greenland S. Sensitivity analysis of misclassification: a graphical and a Bayesian approach. *Annals of Epidemiology*. 2006; 16: 834-41.
12. Davidov O, Faraggi D and Reiser B. Misclassification in logistic regression with discrete covariates. *Biometrical Journal*. 2003; 45: 541-53.
13. Hofler M. The effect of misclassification on the estimation of association: a review. *International Journal of Methods in Psychiatric Research*. 2005; 14: 92-101.
14. Brenner H and Savitz DA. The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. *American Journal of Epidemiology*. 1990; 132: 181-92.
15. Magder LS and Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*. 1997; 146: 195-203.
16. McInturff P, Johnson WO, Cowling D and Gardner IA. Modelling risk when binary outcomes are subject to error. *Statistics in Medicine*. 2004; 23: 1095-109.
17. Greenland S and Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. *International Journal of Epidemiology*. 1983; 12: 93-7.
18. Greenland S. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*. 1996; 25: 1107-16.
19. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*. 1988; 7: 745-57.
20. Lash TL, Fox MP and Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer Science+Business Media, LLC, 2009.

21. Gerlach R and Stamey J. Bayesian model selection for logistic regression with misclassified outcomes. *Statistical Modelling*. 2007; 7: 255-73.
22. Turner EL, Lane JA, Donovan JL, et al. Association of diabetes mellitus with prostate cancer: Nested case-control study (ProtecT: Prostate testing for cancer and treatment). *International Journal of Cancer*. 2011; 128: 440-6.
23. Lyles RH and Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in Medicine*. 2010; 29: 2297-309.
24. Lane JA, Hamdy FC, Martin RM, Turner EL, Neal DE and Donovan JL. Latest results from the UK trials evaluating prostate cancer screening and treatment: The CAP and ProtecT studies. *European Journal of Cancer*. 2010; 46: 3095-101.
25. Cancer Research UK. CancerStats report Prostate Cancer - UK, Cancer Research UK. *http://infocancerresearchukorg/cancerstats/types/prostate*. 2010.
26. Bonovas S, Filioussi K and Tsantes A. Diabetes mellitus and risk of prostate cancer: a meta-analysis. *Diabetologia*. 2004; 47: 1071-8.
27. Kristal AR and Stanford JL. Cruciferous vegetables and prostate cancer risk: confounding by PSA screening. *Cancer Epidemiology Biomarkers Prevention*. 2004; 13: 1265.
28. Werny DM, Saraiya M and Gregg EW. Prostate-specific Antigen Values in Diabetic and Nondiabetic US Men, 2001-2002. *American Journal of Epidemiology*. 2006; 164: 978-83.
29. Muller H, Raum E, Rothenbacher D, Stegmaier C and Brenner H. Association of Diabetes and Body Mass Index with Levels of Prostate-Specific Antigen: Implications for Correction of Prostate-Specific Antigen Cutoff Values? *Cancer Epidemiology Biomarkers Prevention*. 2009; 18: 1350-6.
30. Thompson IM, Ankerst DP, Chi C, et al. Operating Characteristics of Prostate-Specific Antigen in Men With an Initial PSA Level of 3.0 ng/mL or Lower. *JAMA: The Journal of the American Medical Association*. 2005; 294: 66-70.
31. Terris MK. Sensitivity and specificity of sextant biopsies in the detection of prostate cancer: preliminary report. *Urology*. 1999; 54: 486-9.
32. Rocco B, de Cobelli O, Leon ME, et al. Sensitivity and Detection Rate of a 12-Core Trans-Perineal Prostate Biopsy: Preliminary Report. *European Urology*. 2006; 49: 827-33.
33. Porta M, Malats N, Pinol JL, Rifa J, Andreu M and Real FX. Diagnostic certainty and potential for misclassification in exocrine pancreatic cancer. PANKRAS I Project Investigations. *Journal of Clinical Epidemiology*. 1994; 47: 1069-79.
34. Blair A, Stewart P, Lubin JH and Forastiere F. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *American Journal of Industrial Medicine*. 2007; 50: 199-207.
35. Jurek AM, Greenland S and Maldonado G. How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null? *International Journal of Epidemiology*. 2008; 37: 382-5.