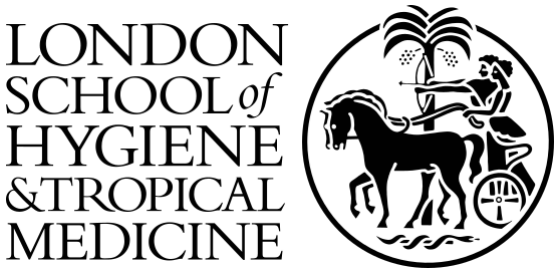

Appendix A – Algorithms

1. Sensitivity/Specificity Imputation Analysis (SS)

A. Fixed-parameter bias-sensitivity analysis (SS FBA)

To begin, we estimated the associations between the misclassified smoking status and the other variables by fitting the following unconditional logistic regression model to the misclassified data:

$$\text{logit}\left[P\left(X=1|Y=y,C=c\right)\right]=\gamma_0+\gamma_Y y+\gamma_C c+\gamma_{YC} yc \qquad (1)$$

We specified initial values Se^0 for sensitivity and Sp^0 for specificity, which were set to combinations of the following values: Se^0 (0.7, 0.8, 0.9) and Sp^0 (0.8, 0.9, 1) (Table A1). For each combination, FBA was applied according to the following steps:

i. Estimation of $\pi^* = P\left(X=1\right)$ for each individual using the fitted model (1) coefficients and their values of *y* and *c*.

ii. Restriction of Se^0 and Sp^0 according to the following equations in order to confine $\pi = P\left(T=1\right)$ to values between 0 and 1

$$Se = \max\left(Se^0, \hat{\pi}^*\right);\ Sp = \max\left(Sp^0, 1-\hat{\pi}^*\right)$$

where $\hat{\pi}^*$ are the estimates of π^* for each individual obtained in i.

iii. Calculation of the positive predictive values (PPV) and the negative predictive values (NPV) according to the following equations

$$PPV = \frac{Se\times\left(\hat{\pi}^* + Sp - 1\right)}{\hat{\pi}^*\times\left(Se + Sp - 1\right)};\ NPV = \frac{Sp\times\left(Se-\hat{\pi}^*\right)}{\left(1-\hat{\pi}^*\right)\left(Se + Sp - 1\right)}$$

iv. Calculation of the frequencies of subjects in each strata $T\times Y\times C$ from PPV and NPV

v. Estimation of the *C*-adjusted ln OR_{TY}

vi. Use of the jackknife procedure[21, 22] to account for the uncertainty in the estimation of $\hat{\pi}^*$ (reiteration of steps i. to v. for each 'leave one out' sample from the original data) and calculation of the interval estimate for ln OR_{TY} from the jackknife standard error (SE).

B. Probabilistic bias-sensitivity analysis (SS PBA)

We assumed that self-reported smoking status was better than chance [18] (i.e. sensitivity and specificity were both greater than 0.5). Logit-transformed scaled normal prior distributions were therefore specified for sensitivity and specificity so that both parameters would fall in the interval [0.5,1]. Specifically, we defined

$$Se^0 = 0.5 + 0.5\text{expit}(\lambda)$$
$$Sp^0 = 0.5 + 0.5\text{expit}(\varepsilon)$$

with normal prior distributions on λ and ε, as specified in Table A1. In order to determine the parameters for these prior distributions, we first chose 95% limits for sensitivity and specificity, converted these limits into limits for λ and ε by solving the above equations, and calculated prior means and prior standard deviations for λ and ε from these limits.

The association between misclassified smoking status and the other variables was then estimated by again fitting model (1) to the data.

PBA was implemented via the following MCSA algorithm:

i. For 10,000 iterations

   a. Estimation of $\pi^* = P(X=1)$ for each individual using the estimated model (1) coefficients

   b. Random draws of Se^0 and Sp^0 from their prior distributions

c. Restriction of Se^0 and Sp^0: $Se = \max\left(Se^0, \hat{\pi}^*\right)$; $Sp = \max\left(Sp^0, 1-\hat{\pi}^*\right)$ where $\hat{\pi}^*$ are the estimates of π^* for each individual obtained in a.

d. Calculation of PPV and NPV for each individual

e. Calculation of the frequencies of subjects in each strata $T \times Y \times C$ from PPV and NPV

f. Estimation of the *C*-adjusted ln OR_{TY}

g. Use of the jackknife procedure to account for the uncertainty in the estimation of π^* (reiteration of steps a. to f. for each 'leave one out' sample from the original data) and calculation of the jackknife standard error of ln OR_{TY}.

h. Perturbation of ln OR_{TY} with its jackknife standard error: $\ln\widetilde{OR}_{TY}$ = random draw from a normal distribution with mean = ln OR_{TY} and standard deviation = jackknife SE(ln OR_{TY})

ii. Computation of the mean, median, and 2.5th and 97.5th percentiles from the distribution of the 10,000 ln OR_{TY} and $\ln\widetilde{OR}_{TY}$ estimates, and their antilogs. We refer to the resulting 2.5th and 97.5th simulation percentiles for OR_{TY} and $\widetilde{OR}_{TY}$ as 95% simulation limits (SL) for OR_{TY}, under the given (possibly truncated) priors.

Table A1 Priors on sensitivity and specificity for SS

SS FBA				
			Values for sensitivity and specificity	
Set of priors			se^0	sp^0
1			0.7	0.8
2			0.7	0.9
3			0.7	1
4			0.8	0.8
5			0.8	0.9
6			0.8	1
7			0.9	0.8
8			0.9	0.9
9			0.9	1
SS PBA				
	Priors parameters mean (standard deviation)		**Means [95% limits] for sensitivity and specificity**	
Set of priors	λ	ε	se^0	sp^0
1	-0.41(0.5)	0.41(0.5)	0.7[0.60,0.82]	0.8[0.68,0.90]
2	0.41(0.5)	1.39(0.5)	0.8[0.68,0.90]	0.9[0.80,0.96]
3	0.41(1.5)	1.39(1.5)	0.8[0.54,0.98]	0.9[0.59,0.99]
4	1.39(0.5)	3.89(0.5)	0.9[0.80,0.96]	0.99[0.97,1.00[

2. Direct Imputation Analysis (DI)

A. Fixed-parameter bias-sensitivity analysis (DI FBA)

The probability of being a 'true' ever-smoker $P(T = 1)$ was estimated from a logistic regression (model (2)).

$$\text{logit}\left[P\left(T=1\middle|X=x,Y=y,C=c\right)\right]=\beta_0+\beta_X x+\beta_Y y+\beta_C c+\beta_{YC} yc \qquad (2)$$

where the *xc* and *xy* product terms are omitted because they are zero under nondifferential misclassification.

We gave fixed values to all model (2) coefficients as shown in Table A2. The values were obtained by translating values for sensitivity, specificity, ln OR_{TY} and prevalences of true smokers in strata of *Y* and *C* that had been chosen on the basis of published data, surveys and our prior assumptions. Details of the calculations are available in Appendix C. Unlike sensitivity and specificity, model (2) coefficients have no logical range restrictions.

The following algorithm was then applied:

i. Computation of $\hat{\pi}$, the estimate of $P(T=1)$ from model (2) and calculation of PPV and NPV for each individual

ii. Calculation of the frequencies of subjects in each strata $T \times Y \times C$ from PPV and NPV

iii. Estimation of the C-adjusted ln OR_{TY} and 95% CI.

B. Probabilistic bias-sensitivity analysis (DI PBA)

The probability of being a 'true' ever-smoker $P(T = 1)$ was represented by model (2).

We then placed normal prior distributions on all model (2) coefficients as shown in Table A2. Means and 95% limits for sensitivity, specificity, ln OR_{TY} and prevalences of true smokers in strata of *Y* and *C* were translated into prior means and standard deviations for model (2) coefficients using the equations in Appendix B. In order to allow the comparison between DI PBA, SS PBA and the fully Bayesian analysis described next, the prior means, standard deviations and correlation for coefficients β_0 and β_X were estimated by simulation (see Appendix D for details). If DI PBA were used alone with no intent to compare it with other methods, parameters for the prior distributions of β_0 and β_X could be specified directly based on background information, without simulation. Appendix E provides an approximate estimate of the correlation ρ_{0X} between β_0 and β_X.

The following MCSA algorithm was then applied:

i. For 100,000 iterations

   a. Random draw of model (2) coefficients from their respective prior distributions

   b. Computation of $\hat{\pi}$, the estimate of $P(T=1)$ from model (2) for each individual

   c. Imputation of *T* from a Bernoulli distribution with probability of success $\hat{\pi}$

   d. Computation of a *C*-adjusted ln OR_{TY} from the imputed *TYC* data

e. Perturbation of ln OR_{TY} with the original *XY* random error: $\ln \widetilde{OR}_{TY}$ = random draw from a normal distribution with mean= ln OR_{TY} and standard deviation = SE(ln OR_{XY})

ii. Computation of the mean, median, and 2.5^{th} and 97.5^{th} percentiles from the distribution of the 100,000 ln OR_{TY} and $\ln \widetilde{OR}_{TY}$ estimates, and their antilogs.

Table A2 Priors on model (2) coefficients for DI

DI FBA												
	Fixed values					Values for sensitivity, specificity, OR_{TY}, OR_{TC}, and prevalence of T=1					
Set of values	β_0	$\beta_{X\ (a)}$	$\beta_{Y\ (a)}$	$\beta_{C\ (a)}$	$\beta_{YC\ (a)}$	sensitivity	specificity	$OR_{TY}(C=0)$	$OR_{TC}(Y=0)$	$OR_{TY}(C=1)/OR_{TY}(C=0)$	P(T=1|Y=0,C=0)
1	-1.37	2.23	1.94	0.10	0.46	0.7	0.8	6.93	1.11	1.59	0.40
2	-1.90	3.58	1.94	0.10	0.46	0.8	0.9	6.93	1.11	1.59	0.40
3	-2.69	6.79	1.94	0.10	0.46	0.9	0.99	6.93	1.11	1.59	0.40
4	-1.37	2.23	1.25	0.10	0.46	0.7	0.8	3.5	1.11	1.59	0.40
5	-1.90	3.58	1.25	0.10	0.46	0.8	0.9	3.5	1.11	1.59	0.40
6	-2.69	6.79	1.25	0.10	0.46	0.9	0.99	3.5	1.11	1.59	0.40
7	-1.37	2.23	2.64	0.10	0.46	0.7	0.8	14	1.11	1.59	0.40
8	-1.90	3.58	2.64	0.10	0.46	0.8	0.9	14	1.11	1.59	0.40
9	-2.69	6.79	2.64	0.10	0.46	0.9	0.99	14	1.11	1.59	0.40

DI PBA												
Set of priors	Priors parameters						Means [95% limits] for sensitivity, specificity, OR_{TY}, OR_{TC}, and prevalence of T=1					
	mean (standard deviation)					Correlation (β_0 , β_X)						
	β_0	$\beta_{X\ (a)}$	$\beta_{Y\ (a)}$	$\beta_{C\ (a)}$	$\beta_{YC\ (a)}$	ρ_{0X}	sensitivity	specificity	$OR_{TY}(C=0)$	$OR_{TC}(Y=0)$	$OR_{TY}(C=1)/OR_{TY}(C=0)$	P(T=1|Y=0,C=0)
1	-1.39 (0.23)	2.28 (0.47)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.80	0.7[0.60,0.82]	0.8[0.68,0.90]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
2	-1.92 (0.31)	3.64 (0.57)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.72	0.8[0.68,0.90]	0.9[0.80,0.96]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
3	-2.08 (0.93)	3.93 (1.66)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.74	0.8[0.54,0.98]	0.9[0.59,0.99]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
4	-2.71 (0.40)	6.81 (0.66)	1.94 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.65	0.9[0.80,0.96]	0.99[0.97,1.00[	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
5	-1.92 (0.31)	3.64 (0.57)	1.25 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.72	0.8[0.68,0.90]	0.9[0.80,0.96]	3.5 [0.89,13.76]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
6	-1.92 (0.31)	3.64 (0.57)	2.64 (0.7)	0.10 (0.35)	0.46 (0.35)	-0.72	0.8[0.68,0.90]	0.9[0.80,0.96]	14[3.55,55.26]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]

3. Fully Bayesian analysis

Prior distributions for Bayesian analysis were chosen to allow direct comparison between the three methods. The model was modified from Chu et al.[18] to include the sex *C*.

We specified prior distributions for two groups of parameters:

a) The sensitivity and specificity, defining the association between the 'true' smoking status T and the misclassified smoking status *X*

As in SS PBA, we defined:

$$Se^0 = 0.5 + 0.5\text{expit}(\lambda)$$
$$Sp^0 = 0.5 + 0.5\text{expit}(\varepsilon)$$

and we placed normal prior distributions on λ and ε.

b) The association of T with case/control status Y and sex C

The prevalence of 'true' smokers in the population was defined as a function of *Y* and *C* (model (3)).

$$\text{logit}\left[P\left(T=1\middle|Y=y,C=c\right)\right]=\alpha_0+\alpha_Y y+\alpha_C c+\alpha_{YC} yc \qquad (3)$$

We placed normal prior distributions on all model (3) coefficients. The values for the parameters of these prior distributions were obtained by giving means and 95% limits to ln OR_{TY} and prevalences of true smokers in strata of *Y* and *C*, and by converting those into prior means and standard deviations for α_0, α_Y, α_C, α_{YC} using the equations in Appendix B.

Unlike model (2) in DI, model (3) did not include the misclassified smoking status *X* as prior distributions were already specified for the association between *T* and *X* in a). Therefore, while DI model (2) coefficients β_0, β_C, β_Y and β_{YC} are functions of *X*, *Y* and *C*, model (3) coefficients α_0, α_Y, α_C, α_{YC} only depend on *Y* and *C*. However, as seen in Table A3, since we are considering only nondifferential misclassification, the associations between *T* and *Y* and between *T* and *C* do not depend on *X* and hence

$$\beta_Y=\ln\left(OR_{TY}\left(X=x,C=0\right)\right)=\ln\left(OR_{TY}\left(C=0\right)\right)=\alpha_Y$$

$$\beta_C=\ln\left(OR_{TC}\left(X=x,Y=0\right)\right)=\ln\left(OR_{TC}\left(Y=0\right)\right)=\alpha_C$$

$$\begin{aligned}\beta_{YC}&=\ln\left(OR_{TY}\left(X=x,C=1\right)\right)-\ln\left(OR_{TY}\left(X=x,C=0\right)\right)\\&=\ln\left(OR_{TY}\left(C=1\right)\right)-\ln\left(OR_{TY}\left(C=0\right)\right)\\&=\alpha_{YC}\end{aligned}$$

Markov Chain Monte Carlo (MCMC) in WinBUGS was used to sample from the posterior distribution. Two Markov chains were run using the block Gibbs sampler with 800,000 iterations following 10,000 discarded for burn-in.

In the first MCMC analysis, in order to allow direct comparison with SS PBA, we placed the same informative prior distributions on λ and ε as in SS PBA (see Table A1) while we placed vague prior distributions on α_0, α_Y, α_C and α_{YC}, as specified in Table A3.

In the second MCMC analysis, in order to allow direct comparison with DI PBA (see Table A2), we placed informative distributions on all parameters as specified in Table A3. The simulation linking the prior distributions for coefficients λ, ε and α_0 to the prior distributions for DI PBA coefficients β_0 and β_X is described in Appendix D. We then placed the same prior distributions on α_Y, α_C, α_{YC} as on β_Y, β_C, β_{YC}, respectively (Table A3).

SAS and WinBUGS codes are available on request.

Table A3 Prior distributions for the Bayesian (MCMC) analyses corresponding to SS PBA (MCMC Analysis 1) and DI PBA (MCMC Analysis 2)

		Prior values mean (standard deviation)						Means [95% limits] for sensitivity, specificity, OR_{TY}, OR_{TC}, and prevalence of T=1					
MCMC Analysis	**Set of priors**	λ	ε	α_0	α_Y	α_C	α_{YC}	**sensitivity**	**specificity**	**OR_{TY}(C=0)**	**OR_{TC}(Y=0)**	**OR_{TY}(C=1)/ OR_{TY}(C=0)**	**P(T=1|Y=0,C=0)**
1	1	-0.41(0.5)	0.41(0.5)	0(2)	0(1.5)	0(4)	0(3)	0.7[0.60,0.82]	0.8[0.68,0.90]	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]
	2	0.41(0.5)	1.39(0.5)	0(2)	0(1.5)	0(4)	0(3)	0.8[0.68,0.90]	0.9[0.80,0.96]	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]
	3	0.41(1.5)	1.39(1.5)	0(2)	0(1.5)	0(4)	0(3)	0.8[0.54,0.98]	0.9[0.59,0.99]	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]
	4	1.39(0.5)	3.89(0.5)	0(2)	0(1.5)	0(4)	0(3)	0.9[0.80,0.96]	0.99[0.97,1.00[	1[0.05,18.92]	1[0.00,2540.21]	1[0,378.4]	0.5[0.02,0.98]
2	1	-0.41(0.5)	0.41(0.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.7[0.60,0.82]	0.8[0.68,0.90]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
	2	0.41(0.5)	1.39(0.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.68,0.90]	0.9[0.80,0.96]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
	3	0.41(1.5)	1.39(1.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.54,0.98]	0.9[0.59,0.99]	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
	4	1.39(0.5)	3.89(0.5)	-0.39(0.07)	1.94(0.7)	0.1(0.35)	0.46 (0.35)	0.9[0.80,0.96]	0.99[0.97,1.00[	6.93[1.76,27.44]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
	5	0.41(0.5)	1.39(0.5)	-0.39(0.07)	1.25(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.68,0.90]	0.9[0.80,0.96]	3.50[0.89,13.76]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]
	6	0.41(0.5)	1.39(0.5)	-0.39(0.07)	2.64(0.7)	0.1(0.35)	0.46 (0.35)	0.8[0.68,0.90]	0.9[0.80,0.96]	14.00[3.55,55.26]	1.11[0.56,2.19]	1.59[0.80,3.15]	0.40[0.37,0.44]

Appendix B – Definition of model (2) (DI) and model (3) (MCMC) coefficients

Table B1 Definition of model (2) coefficients (DI)

Coefficient	Definition
β_0	$\exp it(\beta_0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ $= P(T=1 \mid X=0, Y=0, C=0)$ $= \frac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0) + P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}$ $= \frac{\left(\frac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0)}\right)}{\left(1+\frac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0)}\right)}$ Then, $\exp(\beta_0) = \frac{P(X=0 \mid T=1, Y=0, C=0)P(T=1 \mid Y=0, C=0)}{P(X=0 \mid T=0, Y=0, C=0)P(T=0 \mid Y=0, C=0)}$ $= \frac{(1-Se_{00})P(T=1 \mid Y=0, C=0)}{Sp_{00}(1-P(T=1 \mid Y=0, C=0))}$ $= \frac{(1-Se)P(T=1 \mid Y=0, C=0)}{Sp(1-P(T=1 \mid Y=0, C=0))}$ as we are assuming nondifferentiality

Cont.

Coefficient	Definition
β_X	$\exp(\beta_X) = \dfrac{\left[\dfrac{P(T=1\mid X=1,Y=y,C=c)}{P(T=0\mid X=1,Y=y,C=c)}\right]}{\left[\dfrac{P(T=1\mid X=0,Y=y,C=c)}{P(T=0\mid X=0,Y=y,C=c)}\right]} = \dfrac{\left[\dfrac{P(X=1\mid T=1,Y=y,C=c)}{P(X=0\mid T=1,Y=y,C=c)}\right]}{\left[\dfrac{P(X=1\mid T=0,Y=y,C=c)}{P(X=0\mid T=0,Y=y,C=c)}\right]} = \dfrac{Se_{yc}\times Sp_{yc}}{(1-Se_{yc})\times(1-Sp_{yc})}$ $= \dfrac{Se\times Sp}{(1-Se)\times(1-Sp)}$ as we are assuming nondifferentiality
β_Y	$\exp(\beta_Y) = \mathrm{OR}_{TY}(X=x,C=0) = \mathrm{OR}_{TY}(C=0)$, as we are assuming nondifferentiality
β_C	$\exp(\beta_C) = \dfrac{\dfrac{P(T=1\mid C=1,X=x,Y=0)}{P(T=0\mid C=1,X=x,Y=0)}}{\dfrac{P(T=1\mid C=0,X=x,Y=0)}{P(T=0\mid C=0,X=x,Y=0)}} = OR_{TC}(X=x,Y=0) = OR_{TC}(Y=0)$, as we are assuming nondifferentiality
β_{YC}	$\exp(\beta_{YC}) = \dfrac{\left\{\left[\dfrac{P(T=1\mid X=x,Y=1,C=1)}{P(T=0\mid X=x,Y=1,C=1)}\right]\right\}}{\left\{\left[\dfrac{P(T=1\mid X=x,Y=0,C=1)}{P(T=0\mid X=x,Y=0,C=1)}\right]\right\}} = \dfrac{\left\{\left[\dfrac{P(Y=1\mid T=1,X=x,C=1)}{P(Y=0\mid T=1,X=x,C=1)}\right]\right\}}{\left\{\left[\dfrac{P(Y=1\mid T=0,X=x,C=1)}{P(Y=0\mid T=0,X=x,C=1)}\right]\right\}} = \dfrac{OR_{TY}(X=x,C=1)}{OR_{TY}(X=x,C=0)}$, $= \dfrac{OR_{TY}(C=1)}{OR_{TY}(C=0)}$, as we are assuming nondifferentiality

Table B2 Definition of model (3) coefficients (MCMC)

Coefficient	Definition
α_0	$\text{expit}(\alpha_0) = \dfrac{\exp(\alpha_0)}{1+\exp(\alpha_0)} = P(T=1 \mid Y=0, C=0)$
α_Y	$\exp(\alpha_Y) = \text{OR}_{TY}(C=0)$
α_C	$\exp(\alpha_C) = OR_{TC}(Y=0)$
α_{YC}	$\exp(\alpha_{YC}) = \dfrac{\left\{\left[\dfrac{P(T=1 \mid Y=1, C=1)}{P(T=0 \mid Y=1, C=1)}\right]\right\}}{\left\{\left[\dfrac{P(T=1 \mid Y=0, C=1)}{P(T=0 \mid Y=0, C=1)}\right]\right\}} = \dfrac{\left\{\left[\dfrac{P(Y=1 \mid T=1, C=1)}{P(Y=0 \mid T=1, C=1)}\right]\right\}}{\left\{\left[\dfrac{P(Y=1 \mid T=0, C=1)}{P(Y=0 \mid T=0, C=1)}\right]\right\}} = \dfrac{OR_{TY}(C=1)}{OR_{TY}(C=0)}$,

Appendix C - Details of the calculation of the values for model (2) coefficients in DI FBA

As mentioned in Appendix B, we assume nondifferentiality i.e. $Se_{00}=Se_{01}=Se_{10}=Se_{11}=Se$ and $Sp_{00}=Sp_{01}=Sp_{10}=Sp_{11}=Sp$, where Se_{yc} and Sp_{yc} are the sensitivity and the specificity for $Y=y$ and $C=c$.

According to the 2009 New Zealand Tobacco Use Survey (NZTUS), the prevalence $P(T=1|Y=0,C=0)$ of ever-smokers (current smokers and ex-smokers) among women in New Zealand is 0.403 and the prevalence $P(T=1|Y=0,C=1)$ of ever-smokers among men in New Zealand is 0.428.

Model (2) coefficients β_0, β_X and β_Y are assigned to different sets of values while coefficients β_Y and β_{YC} are set constant.

Calculation of β_0:

$$\beta_0 = \ln\left(\frac{(1-Se)\times P(T=1|Y=0,C=0)}{Sp\times(1-P(T=1|Y=0,C=0))}\right)$$

- Sets of values 1,4,7: $\beta_0 = \ln\left(\frac{0.3\times 0.403}{0.8\times 0.597}\right) = -1.37$
- Sets of values 2,5,8: $\beta_0 = \ln\left(\frac{0.2\times 0.403}{0.9\times 0.597}\right) = -1.90$
- Sets of values 3,6,9: $\beta_0 = \ln\left(\frac{0.1\times 0.403}{0.99\times 0.597}\right) = -2.69$

Calculation of β_X :

$$\beta_X = \ln\left(\frac{Se \times Sp}{(1-Se)\times(1-Sp)}\right)$$

- Sets of values 1,4,7: $\beta_X = \ln\left(\frac{0.7\times 0.8}{0.3\times 0.2}\right) \approx 2.23$
- Sets of values 2,5,8: $\beta_X = \ln\left(\frac{0.8\times 0.9}{0.2\times 0.1}\right) \approx 3.58$
- Set of values 3,6,9: $\beta_X = \ln\left(\frac{0.9\times 0.99}{0.1\times 0.01}\right) \approx 6.79$

Calculation of β_Y:

$$\beta_Y = \ln\left(OR_{TY}(C=0)\right)$$

- Sets of values 1,2,3: $\beta_Y = \ln(6.93) \approx 1.94$ (where 6.93 is the smoking-lung cancer odds ratio for women in our original data)
- Set of values 4,5,6: $\beta_Y = \ln(3.5) \approx 1.25$
- Set of values 7,8,9: $\beta_Y = \ln(14) \approx 2.64$

Calculation of β_C :

$$\beta_C = \ln\left(OR_{TC}(Y=0)\right) = \frac{\frac{P(T=1|C=1,Y=0)}{1-P(T=1|C=1,Y=0)}}{\frac{P(T=1|C=0,Y=0)}{1-P(T=1|C=0,Y=0)}} = \frac{\frac{0.428}{0.572}}{\frac{0.403}{0.597}} \approx 0.10$$

Calculation of β_{YC}:

- $\beta_{YC} = \ln\left(\frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}\right) = \frac{11.03}{6.93} \approx 0.46$ (where 11.03 and 6.93 are the smoking-lung cancer odds ratios for men and women, respectively, in our original data)

Appendix D - Details of the calculation of the prior distribution parameters for model (2) coefficients in DI PBA

Let:

- $\lambda^{mean}, \lambda^{sd}, \varepsilon^{mean}, \varepsilon^{sd}$ be the prior means and standard deviations given to MCMC analysis 2 parameters λ and ε, respectively.
- ${\alpha_0}^{mean}$ and ${\alpha_0}^{sd}$ be the prior mean and standard deviation given to MCMC analysis 2 parameter α_0.
- $OR_{TY}(C=0)^{mean}, OR_{TC}(Y=0)^{mean}, \frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}^{mean}$ be the means given to $\mathrm{OR}_{TY}(C=0)$, $\mathrm{OR}_{TC}(Y=0)$, $\frac{\mathrm{OR}_{TY}(C=1)}{\mathrm{OR}_{TY}(C=0)}$, respectively.
    - $OR_{TY}(C=0)^{lower}, OR_{TC}(Y=0)^{lower}, \frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}^{lower}$ be the lower 95% limits given to $\mathrm{OR}_{TY}(C=0)$, $\mathrm{OR}_{TC}(Y=0)$, $\frac{\mathrm{OR}_{TY}(C=1)}{\mathrm{OR}_{TY}(C=0)}$, respectively.
    - $OR_{TY}(C=0)^{upper}, OR_{TC}(Y=0)^{upper}, \frac{OR_{TY}(C=1)}{OR_{TY}(C=0)}^{upper}$ be the upper 95% limits given to $\mathrm{OR}_{TY}(C=0)$, $\mathrm{OR}_{TC}(Y=0)$, $\frac{\mathrm{OR}_{TY}(C=1)}{\mathrm{OR}_{TY}(C=0)}$, respectively.

Calculation of prior mean $\beta_0{}^{mean}$ and prior standard deviation $\beta_0{}^{sd}$ for β_0 , prior mean $\beta_X{}^{mean}$ and prior standard deviation $\beta_X{}^{sd}$ for β_X , and prior correlation ρ_{0X} between β_0 and β_X

i. For 10,000 iterations

   a. Random draw of λ from $N\left(\lambda^{mean}, \lambda^{sd}\right)$
   b. Random draw of ε from $N\left(\varepsilon^{mean}, \varepsilon^{sd}\right)$
   c. Random draw of α_0 from $N\left(\alpha_0{}^{mean}, \alpha_0{}^{sd}\right)$
   d. Computation of $Se = 0.5 + 0.5\text{expit}(\lambda)$
   e. Computation of $Sp = 0.5 + 0.5\text{expit}(\varepsilon)$
   f. Computation of $\beta_0 = \ln\left(\frac{(1-Se)\times P(T=1|Y=0,C=0)}{Sp\times(1-P(T=1|Y=0,C=0))}\right) = \ln\left(\frac{(1-Se)}{Sp}\right) + \alpha_0$
   g. Computation of $\beta_X = \ln\left(\frac{Se\times Sp}{(1-Se)\times(1-Sp)}\right)$

ii. Computation of $\beta_0{}^{mean}$, $\beta_0{}^{sd}$, $\beta_X{}^{mean}$, $\beta_X{}^{sd}$, ρ_{0X} from the 10,000 values obtained for β_0 and β_X

Calculation of prior mean $\beta_Y{}^{mean}$ and prior standard deviation $\beta_Y{}^{sd}$ for β_Y

$$\beta_Y{}^{mean} = \ln\left(OR_{TY}(C=0)^{mean}\right)$$

$$\beta_Y{}^{sd} = \frac{\beta_Y{}^{upper} - \beta_Y{}^{lower}}{2\times 1.96},$$

where $\beta_Y{}^{lower} = \ln\left(OR_{TY}(C=0)^{lower}\right)$ and $\beta_Y{}^{upper} = \ln\left(OR_{TY}(C=0)^{upper}\right)$

Calculation of prior mean β_C^{mean} and prior standard deviation β_C^{sd} for β_C

$$\beta_C^{mean} = \ln\left(OR_{TC}(Y=0)^{mean}\right)$$

$$\beta_C^{sd} = \frac{\beta_C^{upper} - \beta_C^{lower}}{2 \times 1.96},$$

where $\beta_C^{lower} = \ln\left(OR_{TC}(Y=0)^{lower}\right)$ and $\beta_C^{upper} = \ln\left(OR_{TC}(Y=0)^{upper}\right)$

Calculation of prior mean β_{YC}^{mean} and prior standard deviation β_{YC}^{sd} for β_{YC}

$$\beta_{YC}^{mean} = \ln\left(\frac{OR_{TY}(C=1)^{mean}}{OR_{TY}(C=0)}\right)$$

$$\beta_{YC}^{sd} = \frac{\beta_{YC}^{upper} - \beta_{YC}^{lower}}{2 \times 1.96},$$

where $\beta_{YC}^{lower} = \ln\left(\frac{OR_{TY}(C=1)^{lower}}{OR_{TY}(C=0)}\right)$ and $\beta_{YC}^{upper} = \ln\left(\frac{OR_{TY}(C=1)^{upper}}{OR_{TY}(C=0)}\right)$

Appendix E – Approximate estimate of the correlation ρ_{0X} between β_0 and β_X

In model (2), β_Y, β_C and β_{YC} are pre-specified and *y*, *c* and *yc* are known.

Let $k = \beta_Y y + \beta_C c + \beta_{YC} yc$.

Then, model (2) becomes:

$$\text{logit}\left[P\left(T=1 \middle| X=x, Y=y, C=c\right)\right] = k + \beta_0 + \beta_X x$$

Using a linear approximation $u = \left(k + \beta_0\right) + \beta_X x$, which is linear in x,

we obtain from standard results for linear regression:

$$\rho_{0X} \approx -\frac{\sum_{i=1}^{N} x_i}{\sqrt{\sum_{i=1}^{N} {x_i}^2 N}}$$