Corbin, M; Haslett, S; Pearce, N; Maule, M; Greenland, S; (2017) A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable. International journal of epidemiology. ISSN 0300-5771
DOI: https://doi.org/10.1093/ije/dyx027

Downloaded from: http://researchonline.lshtm.ac.uk/id/eprint/3682730/

DOI: https://doi.org/10.1093/ije/dyx027

**Usage Guidelines:**

Please refer to usage guidelines at https://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/
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:

\[
\logit \left[ P(X = 1|Y = y, C = c) \right] = \gamma_0 + \gamma_y y + \gamma_c c + \gamma_{yc} yc
\]  

(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 \(\pi^* = P(X = 1)\) 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 \(\pi = P(T = 1)\) to values between 0 and 1

\[
Se = \max(Se^0, \hat{\pi}^*); \quad Sp = \max(Sp^0, 1 - \hat{\pi}^*)
\]

where \(\hat{\pi}^*\) are the estimates of \(\pi^*\) 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

\[
PPV = \frac{Se \times (\hat{\pi}^* + Sp - 1)}{\hat{\pi}^* \times (Se + Sp - 1)}; \quad NPV = \frac{Sp \times (Se - \hat{\pi}^*)}{(1 - \hat{\pi}^*) (Se + Sp - 1)}
\]

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 $\ln \text{OR}_{TY}$ 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

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

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

with normal prior distributions on $\lambda$ and $\varepsilon$, 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 $\lambda$ and $\varepsilon$ by solving the above equations, and calculated prior means and prior standard deviations for $\lambda$ and $\varepsilon$ 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 $Sp^0$: 

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

where $\hat{\pi}^*$ are the estimates of $\pi^*$ 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 $\pi^*$ (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: 

$$\ln \tilde{OR}_{TY} = \text{random draw from a normal distribution with mean} = \ln OR_{TY} \text{ and standard deviation}$$

$$= \text{jackknife SE(} \ln OR_{TY})$$

ii. Computation of the mean, median, and 2.5\textsuperscript{th} and 97.5\textsuperscript{th} percentiles from the distribution of the 10,000 $\ln OR_{TY}$ and $\ln \tilde{OR}_{TY}$ estimates, and their antilogs. We refer to the resulting 2.5\textsuperscript{th} and 97.5\textsuperscript{th} simulation percentiles for $OR_{TY}$ and $\tilde{OR}_{TY}$ as 95% simulation limits (SL) for $OR_{TY}$, under the given (possibly truncated) priors.
Table A1 Priors on sensitivity and specificity for SS

<table>
<thead>
<tr>
<th>Set of priors</th>
<th>ss0</th>
<th>sp0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>6</td>
<td>0.8</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>9</td>
<td>0.9</td>
<td>1</td>
</tr>
</tbody>
</table>

SS PBA

<table>
<thead>
<tr>
<th>Set of priors</th>
<th>Priors parameters mean (standard deviation)</th>
<th>Means [95% limits] for sensitivity and specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \lambda )</td>
<td>( \varepsilon )</td>
</tr>
<tr>
<td>1</td>
<td>-0.41(0.5)</td>
<td>0.41(0.5)</td>
</tr>
<tr>
<td>2</td>
<td>0.41(0.5)</td>
<td>1.39(0.5)</td>
</tr>
<tr>
<td>3</td>
<td>0.41(1.5)</td>
<td>1.39(1.5)</td>
</tr>
<tr>
<td>4</td>
<td>1.39(0.5)</td>
<td>3.89(0.5)</td>
</tr>
</tbody>
</table>

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)).

\[
\logit \left[ P(T = 1|X, Y, C = e) \right] = \beta_0 + \beta_x x + \beta_y y + \beta_{cC} c + \beta_{yc} yc \quad (2)
\]

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 \( \ln \text{OR}_{TY} \) 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, \( \ln \text{OR}_{TY} \) 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 \( \beta_0 \) and \( \beta_x \) 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 \( \beta_0 \) and \( \beta_x \) could be specified directly based on background information, without simulation. Appendix E provides an approximate estimate of the correlation \( \rho_{0x} \) between \( \beta_0 \) and \( \beta_x \).

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 \( \ln \text{OR}_{TY} \) from the imputed \( TYC \) data
e. Perturbation of ln OR\textsubscript{TY} with the original XY random error: ln\tilde{OR}_{TY} = \text{random draw from a normal distribution with mean= ln OR\textsubscript{TY} and standard deviation = SE(ln OR\textsubscript{XY})}

ii. Computation of the mean, median, and 2.5\textsuperscript{th} and 97.5\textsuperscript{th} percentiles from the distribution of the 100,000 ln OR\textsubscript{TY} and ln\tilde{OR}_{TY} estimates, and their antilogs.
Table A2 Priors on model (2) coefficients for DI

| Set of values | $\beta_0$ | $\beta_X$ (a) | $\beta_Y$ (a) | $\beta_C$ (a) | $\beta_{YC}$ (a) | Fixed values | DI FBA | Values for sensitivity, specificity, OR$_{TC}$, OR$_{TY}$, and prevalence of T=1 | P(T=1|Y=0,C=0) |
|---------------|----------|---------------|---------------|---------------|----------------|--------------|--------|-----------------------------------------------------------------|----------------|
| 1             | -1.37    | 2.23          | 1.94          | 0.10          | 0.46           |              |        | sensitivity 0.7, specificity 0.8, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 2             | -1.90    | 3.58          | 1.94          | 0.10          | 0.46           |              |        | sensitivity 0.8, specificity 0.9, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 3             | -2.69    | 6.79          | 1.94          | 0.10          | 0.46           |              |        | sensitivity 0.9, specificity 0.99, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 4             | -1.37    | 2.23          | 1.25          | 0.10          | 0.46           |              |        | sensitivity 0.7, specificity 0.8, OR$_{TY}(C=0)$ 3.5, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 5             | -1.90    | 3.58          | 1.25          | 0.10          | 0.46           |              |        | sensitivity 0.8, specificity 0.9, OR$_{TY}(C=0)$ 3.5, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 6             | -2.69    | 6.79          | 1.25          | 0.10          | 0.46           |              |        | sensitivity 0.9, specificity 0.99, OR$_{TY}(C=0)$ 3.5, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 7             | -1.37    | 2.23          | 2.64          | 0.10          | 0.46           |              |        | sensitivity 0.7, specificity 0.8, OR$_{TY}(C=0)$ 14, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 8             | -1.90    | 3.58          | 2.64          | 0.10          | 0.46           |              |        | sensitivity 0.8, specificity 0.9, OR$_{TY}(C=0)$ 14, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 9             | -2.69    | 6.79          | 2.64          | 0.10          | 0.46           |              |        | sensitivity 0.7, specificity 0.8, OR$_{TY}(C=0)$ 14, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |

**DI PBA**

| Set of priors | Priors parameters | Correlation ($\beta_0$, $\beta_X$) | Means [95% limits] for sensitivity, specificity, OR$_{TC}$, OR$_{TY}$, and prevalence of T=1 | P(T=1|Y=0,C=0) |
|---------------|--------------------|-------------------------------------|------------------------------------------------------------------------------------------------|----------------|
| 1             | -1.39 (0.23)       | 2.28 (0.47)                         | sensitivity 0.7, specificity 0.8, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 2             | -1.92 (0.31)       | 3.64 (0.57)                         | sensitivity 0.8, specificity 0.9, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 3             | -2.08 (0.93)       | 3.93 (1.66)                         | sensitivity 0.8, specificity 0.99, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 4             | -2.71 (0.40)       | 6.81 (0.66)                         | sensitivity 0.9, specificity 0.99, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 5             | -1.92 (0.31)       | 3.64 (0.57)                         | sensitivity 0.8, specificity 0.99, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
| 6             | -1.92 (0.31)       | 3.64 (0.57)                         | sensitivity 0.8, specificity 0.99, OR$_{TY}(C=0)$ 6.93, OR$_{TC}(Y=0)$ 1.11, OR$_{TY}(C=1)/OR_{TY}(C=0)$ 1.59, P(T=1|Y=0,C=0) 0.40 |               |
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:

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

and we placed normal prior distributions on $\lambda$ and $\varepsilon$. 
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)).

$$\logit\left[P(T = 1|Y = y, C = c)\right] = \alpha_0 + \alpha_y y + \alpha_c c + \alpha_{yc} yc$$ (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 $\alpha_0, \alpha_y, \alpha_c, \alpha_{yc}$ 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 $\beta_0, \beta_c, \beta_y$ and $\beta_{yc}$ are functions of $X, Y$ and $C$, model (3) coefficients $\alpha_0, \alpha_y, \alpha_c, \alpha_{yc}$ 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

$$\beta_y = \ln(OR_{TY}(X = x, C = 0)) = \ln(OR_{TY}(C = 0)) = \alpha_y$$

$$\beta_c = \ln(OR_{YC}(X = x, Y = 0)) = \ln(OR_{YC}(Y = 0)) = \alpha_c$$

$$\beta_{yc} = \ln(OR_{TY}(X = x, C = 1)) - \ln(OR_{TY}(X = x, C = 0))$$

$$= \ln(OR_{TY}(C = 1)) - \ln(OR_{TY}(C = 0))$$

$$= \alpha_{yc}$$

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 $\lambda$ and $\varepsilon$ as in SS PBA (see Table A1) while we placed vague prior distributions on $\alpha_0$, $\alpha_r$, $\alpha_c$ and $\alpha_{yc}$, 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 $\lambda$, $\varepsilon$ and $\alpha_0$ to the prior distributions for DI PBA coefficients $\beta_0$ and $\beta_X$ is described in Appendix D. We then placed the same prior distributions on $\alpha_y$, $\alpha_c$, $\alpha_{yc}$ as on $\beta_y$, $\beta_c$, $\beta_{yc}$, 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)

| MCMC Analysis | Set of priors | \( \lambda \) | \( \varepsilon \) | \( \alpha_0 \) | \( \alpha_Y \) | \( \alpha_C \) | \( \alpha_{YC} \) | 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             | 2             | 0.41(1.5)   | 1.39(1.5)   | 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             | 3             | 1.39(0.5)   | 3.89(0.5)   | 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             | 4             | 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] |
| 5             | 5             | 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[0.54,0.98] | 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] |
| 6             | 6             | 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] | 0.9[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] |
Appendix B – Definition of model (2) (DI) and model (3) (MCMC) coefficients

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

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \exp(\beta_0) )</td>
<td>( \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} )</td>
</tr>
<tr>
<td></td>
<td>( = \frac{P(T = 1</td>
</tr>
<tr>
<td></td>
<td>( = \frac{P(X = 0</td>
</tr>
<tr>
<td></td>
<td>( = \frac{1 + P(X = 0</td>
</tr>
</tbody>
</table>

Then,

\( \exp(\beta_0) = \frac{P(X = 0|T = 1, Y = 0, C = 0)P(T = 1|Y = 0, C = 0)}{P(X = 0|T = 0, Y = 0, C = 0)P(T = 0|Y = 0, C = 0)} \)

\( = \frac{(1 - S_{e00})P(T = 1|Y = 0, C = 0)}{S_{p00}(1 - P(T = 1|Y = 0, C = 0))} \)

\( = \frac{(1 - S_{e})P(T = 1|Y = 0, C = 0)}{S_{p}(1 - P(T = 1|Y = 0, C = 0))} \)

as we are assuming nondifferentiality.
<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Definition</th>
</tr>
</thead>
</table>
| $\beta_x$ | \[
\exp(\beta_x) = \frac{P(T = 1|X = 1, Y = y, C = c)}{P(T = 0|X = 1, Y = y, C = c)} = \frac{P(X = 1|T = 1, Y = y, C = c)}{P(X = 0|T = 1, Y = y, C = c)} = \frac{Se_y \times Sp_y}{(1 - Se_y) \times (1 - Sp_y)}
\]

as we are assuming nondifferentiality

| $\beta_x$ | $\exp(\beta_y) = \text{OR}_{TY}(X = x, C = 0) = \text{OR}_{TY}(Y = 0), \text{as we are assuming nondifferentiality}$ |
| $\beta_c$ | \[
\exp(\beta_c) = \frac{P[T = 1|C = 1, X = x, Y = 0]}{P[T = 0|C = 1, X = x, Y = 0]} = \frac{P(X = 1|T = 1, Y = 0)}{P(X = 0|T = 1, Y = 0)} = \text{OR}_{TC}(X = x, Y = 0) = \text{OR}_{TC}(Y = 0), \text{as we are assuming nondifferentiality}
\]

$\beta_{yc}$ | \[
\exp(\beta_{yc}) = \frac{\left[\frac{P[T = 1|X = x, Y = 1, C = 0]}{P[T = 0|X = x, Y = 1, C = 0]}\right]}{\left[\frac{P[Y = 1|T = 1, X = x, C = 0]}{P[Y = 0|T = 1, X = x, C = 0]}\right]} = \frac{\left[\frac{P[Y = 1|T = 0, X = x, C = 0]}{P[Y = 0|T = 0, X = x, C = 0]}\right]}{\left[\frac{P[Y = 1|T = 0, X = x, C = 0]}{P[Y = 0|T = 0, X = x, C = 0]}\right]} = \text{OR}_{TY}(X = x, C = 1) = \text{OR}_{TY}(X = x, C = 0), \text{as we are assuming nondifferentiality}
\]
**Table B2** Definition of model (3) coefficients (MCMC)

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>$\text{expit}(\alpha_0) = \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0)} = P(T = 1</td>
</tr>
<tr>
<td>$\alpha_Y$</td>
<td>$\exp(\alpha_Y) = \text{OR}_{TY}(C = 0)$</td>
</tr>
<tr>
<td>$\alpha_C$</td>
<td>$\exp(\alpha_C) = \text{OR}_{TC}(Y = 0)$</td>
</tr>
</tbody>
</table>
| $\alpha_{YC}$ | \[\text{exp}(\alpha_{YC}) = \frac{\begin{bmatrix} P(T = 1|Y = 1, C = 1) \\
\; P(T = 0|Y = 1, C = 1) \\
\; P(T = 1|Y = 0, C = 1) \\
\; P(T = 0|Y = 0, C = 1) \end{bmatrix}}{\begin{bmatrix} P(T = 0|Y = 1, C = 1) \\
\; P(T = 0|Y = 0, C = 1) \end{bmatrix}} = \frac{\text{OR}_{TY}(C = 1)}{\text{OR}_{TY}(C = 0)}\] |
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. $Se_{00}=Se_{01}=Se_{10}=Se_{11}$ 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 $\beta_0$, $\beta_x$ and $\beta_i$ are assigned to different sets of values while coefficients $\beta_i$ and $\beta_{yc}$ are set constant.

**Calculation of $\beta_0$:**

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

- **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 $\beta_x$:

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

- 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 $\beta_Y$:

$$\beta_Y = \ln(OR_{YY}(C = 0))$$

- 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 $\beta_C$:

$$\beta_C = \ln(OR_{YC}(Y = 0)) = \frac{\frac{P(T = 1|C = 1,Y = 0)}{1 - P(T = 1|C = 1,Y = 0)}}{\frac{P(T = 1|C = 0,Y = 0)}{1 - P(T = 1|C = 0,Y = 0)}} = \frac{0.428}{0.572} \approx 0.10$$
Calculation of $\beta_{yc}$:

- $\beta_{yc} = \ln \left( \frac{OR_{ty}(C = 1)}{OR_{ty}(C = 0)} \right) = \frac{11.03}{6.93} \approx 0.46$ (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:

- $\lambda^{\text{mean}}$, $\lambda^{sd}$, $\epsilon^{\text{mean}}$, $\epsilon^{sd}$ be the prior means and standard deviations given to MCMC analysis 2 parameters $\lambda$ and $\epsilon$, respectively.

- $\alpha_0^{\text{mean}}$ and $\alpha_0^{sd}$ be the prior mean and standard deviation given to MCMC analysis 2 parameter $\alpha_0$.

- $OR_{TY}(C = 0)^{\text{mean}}, OR_{TC}(Y = 0)^{\text{mean}}, \frac{OR_{TY}(C = 1)^{\text{mean}}}{OR_{TY}(C = 0)^{\text{mean}}}$ be the means given to $OR_{TY}(C = 0)$, $OR_{TC}(Y = 0)$, $\frac{OR_{TY}(C = 1)}{OR_{TY}(C = 0)}$, respectively.

- $OR_{TY}(C = 0)^{\text{lower}}, OR_{TC}(Y = 0)^{\text{lower}}, \frac{OR_{TY}(C = 1)^{\text{lower}}}{OR_{TY}(C = 0)^{\text{lower}}}$ be the lower 95% limits given to $OR_{TY}(C = 0)$, $OR_{TC}(Y = 0)$, $\frac{OR_{TY}(C = 1)}{OR_{TY}(C = 0)}$, respectively.

- $OR_{TY}(C = 0)^{\text{upper}}, OR_{TC}(Y = 0)^{\text{upper}}, \frac{OR_{TY}(C = 1)^{\text{upper}}}{OR_{TY}(C = 0)^{\text{upper}}}$ be the upper 95% limits given to $OR_{TY}(C = 0)$, $OR_{TC}(Y = 0)$, $\frac{OR_{TY}(C = 1)}{OR_{TY}(C = 0)}$, respectively.
Calculation of prior mean $\beta_0^{\text{mean}}$ and prior standard deviation $\beta_0^{sd}$ for $\beta_0$, prior mean $\beta_X^{\text{mean}}$ and prior standard deviation $\beta_X^{sd}$ for $\beta_X$, and prior correlation $\rho_{0X}$ between $\beta_0$ and $\beta_X$

i. For 10,000 iterations

a. Random draw of $\lambda$ from $N(\lambda^{mean}, \lambda^{sd})$

b. Random draw of $\epsilon$ from $N(\epsilon^{mean}, \epsilon^{sd})$

c. Random draw of $\alpha_0$ from $N(\alpha_0^{mean}, \alpha_0^{sd})$

d. Computation of $Se = 0.5 + 0.5 \text{expit}(\lambda)$

e. Computation of $Sp = 0.5 + 0.5 \text{expit}(\epsilon)$

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 $\beta_0^{mean}$, $\beta_0^{sd}$, $\beta_X^{mean}$, $\beta_X^{sd}$, $\rho_{0X}$ from the 10,000 values obtained for $\beta_0$ and $\beta_X$

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

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

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

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 $\beta_{C}^{\text{mean}}$ and prior standard deviation $\beta_{C}^{sd}$ for $\beta_{C}$

$$\beta_{C}^{\text{mean}} = \ln\left(\text{OR}_{YC}(Y = 0)^{\text{mean}}\right)$$

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

where $\beta_{C}^{\text{lower}} = \ln\left(\text{OR}_{YC}(Y = 0)^{\text{lower}}\right)$ and $\beta_{C}^{\text{upper}} = \ln\left(\text{OR}_{YC}(Y = 0)^{\text{upper}}\right)$

Calculation of prior mean $\beta_{YC}^{\text{mean}}$ and prior standard deviation $\beta_{YC}^{sd}$ for $\beta_{YC}$

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

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

where $\beta_{YC}^{\text{lower}} = \ln\left(\frac{\text{OR}_{TY}(C = 1)^{\text{lower}}}{\text{OR}_{TY}(C = 0)}\right)$ and $\beta_{YC}^{\text{upper}} = \ln\left(\frac{\text{OR}_{TY}(C = 1)^{\text{upper}}}{\text{OR}_{TY}(C = 0)}\right)$
Appendix E – Approximate estimate of the correlation $\rho_{0X}$ between $\beta_0$ and $\beta_X$

In model (2), $\beta_Y, \beta_C$ and $\beta_{YC}$ 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:

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

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

we obtain from standard results for linear regression:

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