


# Methods to estimate proportion and number of nonexposed cases in a population

Heiko Becher<sup>1,2,3</sup>  | Annette Aigner<sup>2,3</sup>

<sup>1</sup> Universitätsklinikum  
Hamburg-Eppendorf, Institut für  
Medizinische Biometrie und  
Epidemiologie, Hamburg,  
Germany

<sup>2</sup> Charité – Universitätsmedizin Berlin,  
Institut für Biometrie und klinische  
Epidemiologie, Berlin, Germany

<sup>3</sup> Berlin Institute of Health (BIH), Berlin,  
Germany

## Correspondence

Heiko Becher, Institut für Medizinische  
Biometrie und Epidemiologie, Univer-  
sitätsklinikum Hamburg-Eppendorf,  
Martinstr. 52, Hamburg, 20246 Germany.  
Email: [h.becher@uke.de](mailto:h.becher@uke.de)



This article has earned an open data badge  
“**Reproducible Research**” for making  
publicly available the code necessary to  
reproduce the reported results. The results  
reported in this article could fully be repro-  
duced.

## Abstract

National mortality statistics commonly provide disease-specific absolute and relative frequencies of death by sex and age, but not by exposure status. However, it is often of interest to know how many of the diseased individuals, that is the cases, were exposed or not exposed to a specific risk factor. We present two methods to estimate the proportion and the number of exposed and nonexposed cases, both of which require an estimate of the exposure prevalence in the nondiseased population. Method I additionally requires an estimate of the relative effect of exposure, that is a relative risk function if the exposure has a continuous distribution, or a relative risk estimate for each category if the exposure is categorical. Method II additionally requires an estimate of the disease rate among the nonexposed. We provide theoretical justifications, discuss practical limitations, and provide an R script to calculate the probability for nonexposure among the diseased, and compare the approaches. Both methods are subsequently applied to the estimation of the number of never smokers among lung cancer deaths. The two suggested methods rely on the availability of specific data sources and might therefore be applicable in different research settings. Both methods yield unbiased estimates of the number of nonexposed cases, given that the respective underlying assumptions are fulfilled.

## KEYWORDS

dose-response, epidemiology, semi-continuous variable, spike at zero

## 1 | INTRODUCTION

Despite inherent limitations, mortality statistics in industrialized countries, such as Germany, can be considered sufficiently reliable for use in public health or health policy making, for example, to plan interventions (Schelhase & Weber, 2007). However, mortality statistics are based on death certificates, which do not contain information on disease risk factors other than sex and age—as, for example, provided by the German Federal Statistical Office or the WHO (Statistisches Bundesamt, 2019; WHO, 2019). Attempts to include information on smoking habits in death certificates have been investigated and have been shown to be considerably difficult to implement (Sitas et al., 2019). This information would, however,

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

be necessary to directly estimate the proportion of diseased individuals, that is the cases, who were exposed or not exposed to a specific risk factor. This number is of public health interest in order to assess the necessity of future or the effectiveness of previous public health interventions. For example regarding second-hand smoke prevention, the aim is to estimate the number of nonsmokers whose deaths resulting from lung cancer can be attributed to exposure to second-hand smoke, as shown in previous studies, for example, in Heuschmann, Heidrich, Wellmann, Kraywinkel, and Keil (2007). Based on mortality statistics, the yearly number of lung cancer deaths is relatively well known, likewise its distribution among the respective sex and age groups. However, there is no information available on the number of smokers or nonsmokers among these cases. The estimation of the proportion of smokers among cases is needed, but generally not readily available.

Mittleman (1995) uses a similar idea based on Bayes' formula to solve a related problem, which is the estimation of exposure prevalence in a population at risk, where exposure is assumed binary. Using the delta method, Mittleman also gives a variance estimate for the resulting prevalence estimate.

The present paper deals with two methods to estimate the proportion and the total number of cases who were not exposed to a specific risk factor, as this problem has received little attention in the past. In a general scenario, the exposure is zero for nonexposed and follows some continuous distribution among the exposed. Therefore, the exposure is a semicontinuous variable, also called a variable with a spike at zero (Royston, Sauerbrei, & Becher, 2010). Special scenarios are that the exposure is a categorical or a binary variable only, where the latter simply distinguishes exposed from nonexposed. This paper is structured as follows: We present a general approach to estimate the proportion of nonexposed among the cases, given the distribution of the exposure variable in the nondiseased population. One method additionally uses the dose–response relationship, whereas the second is based on an estimate of the disease rate among the nonexposed. Finally, we use a data example of second-hand smoke exposure and lung cancer deaths to demonstrate and compare these methods. We provide an R script to estimate the proportion and absolute number of nonexposed among the cases given some numerical examples and to reproduce the data example.

## 2 | STATISTICAL METHODS

### 2.1 | Preliminaries and notation

We consider mortality or incidence of an outcome  $Y$  (e.g., disease or disease group) with  $Y = 0$  for nondiseased and  $Y = 1$  for diseased and assume that data on the yearly number  $d$  of incident cases or deaths are available by sex  $j$  ( $j = 1, 2$ ) and age group  $k$  ( $k = 1, \dots, K$ ),  $d_{jk}$ , such that the total number of cases is  $d = \sum d_{jk}$ . We further assume that population figures by sex and age group,  $n_{jk}$ , are available, for example, from population registries, then the total population size is  $n = \sum n_{jk}$ . Using the rare disease assumption, we consider these population figures as proxies for the respective figures of the nondiseased population. The numbers for nonexposed are denoted with the subscript 0, that is  $d_0$  denotes the number of nonexposed cases, where for the ease of presentation we generally omit the index for age and sex in the following.

### 2.2 | Method I

#### 2.2.1 | General case

Let  $X$  be the risk factor of interest. We assume that  $X$  takes the value 0 for nonexposed and an arbitrary distribution for the exposed. If the distribution of the exposed is continuous, this is a so-called semicontinuous variable or a variable with a spike at zero (Royston et al., 2010). Generally, the density function  $\tilde{f}_X(x)$  of such a variable is given by

$$\tilde{f}_X(x) = \begin{cases} p_0 & X = 0 \\ (1 - p_0) \cdot f_X(x) & X > 0 \end{cases} \quad (1)$$

with  $p_0 = P(X = 0)$  being the probability of no exposure and  $f_X(x) = \tilde{f}_{X|X>0}(x)$  being the conditional density function for  $X$  given  $X > 0$ . Typical examples for variables with a spike at zero are smoking and alcohol consumption, where a certain proportion of the population has zero exposure, and the nonzero exposure follows some arbitrary continuous distribution.

In a previous paper, the odds ratio (*OR*) function given the distribution of  $X$  in diseased and nondiseased was derived (Becher, Lorenz, Royston, & Sauerbrei, 2012) and can be obtained with the following equation:

$$OR_{X=x \text{ vs } X=x_0} = \frac{\tilde{f}_{X|Y=1}(x) \cdot \tilde{f}_{X|Y=0}(x_0)}{\tilde{f}_{X|Y=0}(x) \cdot \tilde{f}_{X|Y=1}(x_0)}. \quad (2)$$

In (2)  $\tilde{f}_{X|Y=i}(x)$  denotes the conditional density of the spike at zero variable  $X$  given  $Y = i$ , with  $i = 0, 1$ . For  $x_0 = 0$  this simplifies to

$$OR(x) := OR_{X=x \text{ vs } X=0} = \frac{f_{X|Y=1}(x) \cdot (1 - p_{01}) \cdot p_{00}}{f_{X|Y=0}(x) \cdot (1 - p_{00}) \cdot p_{01}}, \quad (3)$$

where  $p_{00}$  and  $p_{01}$  denote the probability of zero exposure in nondiseased and diseased, respectively. Given some distributional assumption, the *OR* function can be derived. For example, if  $f_{X|Y=i}$  is a log-normal distribution with parameters  $(\mu_i, \sigma)$ , the resulting *OR* function is  $OR(x) = \exp(\beta_0 + \beta_1 \log(x))$ , with

$$\beta_0 = \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \ln \left( \frac{(1 - p_{01}) p_{00}}{(1 - p_{00}) p_{01}} \right)$$

and  $\beta_1 = (\mu_1 - \mu_0)/\sigma^2$ . (Becher et al., 2012).

The current setting is different. Here, the distribution of  $X$  in the population and the *OR* function are assumed to be known. Under the rare disease assumption, the distribution of  $X$  in the population can be used as a sufficient approximation of the distribution among the nondiseased. Then we are interested in the distribution of  $X$  among the diseased, particularly in  $p_{01}$ . From (3), we get

$$\frac{f_{X|Y=1}(x) \cdot (1 - p_{01})}{p_{01}} = \frac{OR(x) \cdot f_{X|Y=0}(x) \cdot (1 - p_{00})}{p_{00}},$$

which yields

$$p_{01} = \frac{1}{1 + C} \quad \text{with} \quad C = OR(x) \cdot \frac{f_{X|Y=0}(x) \cdot (1 - p_{00})}{f_{X|Y=1}(x) \cdot p_{00}}. \quad (4)$$

In the following, we derive  $p_{01}$  for selected distributions.

## 2.2.2 | Categorical case

Consider  $X$  being a categorical variable. In the simplest case,  $X$  is binary with values 0 and 1. More generally,  $X$  can take values 0 to  $M$  with probabilities in the nondiseased population of  $p_{00}, p_{10}, p_{20}, \dots, p_{M0}$ . A categorical  $X$  can also be the result of a continuous  $X$  having been categorized into discrete levels, for example, none, low, medium, and high, which is common in epidemiology. With  $OR_{X=m \text{ vs } X=0}$  denoted as  $OR_m$ , the density function is

$$\tilde{f}_{X|Y=0}(x) = \begin{cases} p_{00} & X = 0 \\ p_{10} & X = 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ p_{M0} & X = M \end{cases}, \quad (5)$$

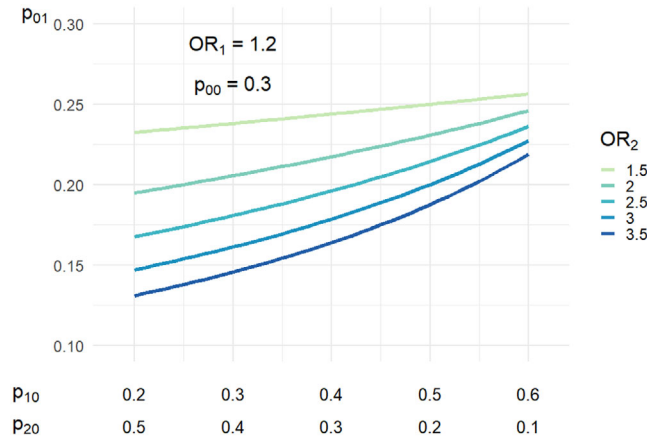


FIGURE 1 Nonexposure probability in cases ( $p_{01}$ ) for  $M = 2$ ,  $OR_1 = 1.2$ ,  $p_{00} = 0.3$ , and some values of  $OR_2$  ( $x$ -axis gives both  $p_{10}$  and  $p_{20} = 1 - p_{00} - p_{10}$ )

TABLE 1 Proportion of nonexposed cases ( $p_{01}$ ) for some selected  $OR$  functions and exposure distribution among the nondiseased

N	OR function: $OR(X = x \text{ vs. } X = 0)$	Distribution of the exposure in the nondiseased: $f_{X Y=0}(x)$	Result for $p_{01}$
1	$OR(X = 1 \text{ vs. } X = 0) = 1.5$ $OR(X = 2 \text{ vs. } X = 0) = 2.5$	Categorical (three categories) $p_{00} = 0.3$ $p_{01} = 0.5$ $p_{02} = 0.2$	0.194
2	$OR: \exp(0.05 \cdot x)$	$0.1x \in (0, 10]$ $0x > 10$ exponential, $\lambda = 0.2$	0.248 0.243
3	$OR: \exp(0.5 \cdot \ln(x))$	$0.1x \in (0, 10]$ $0x > 10$ log-normal, $\mu = \ln(5) - 0.5; \sigma = 1$	0.169 0.178
4	$OR: 2$ Note: constant risk (risk is independent of dose)	$0.1x \in (0, 10]$ $0x > 10$	0.176
5	$OR: \exp(0.3 \cdot \log(x + 1))$	log-normal, $\mu = \ln(5) - 0.5; \sigma = 1$	0.260
6	$OR: \exp(0.1 \cdot (\log(10 \cdot x)) + 0.1)$	log-normal, $\mu = \ln(5) - 0.5; \sigma = 1$	0.250
7	$OR: \exp(0.5 \cdot \ln(x + 1))$	exponential, $\lambda = 0.2$	0.158

1:  $p_{00} = 0.3$  and  $X$  categorical (three categories 0,1,2).

2-7:  $p_{00} = 0.3$  and  $X|X > 0$  continuous with  $E(X|X > 0) = 5$ .

and the solution for  $p_{01}$  is given as

$$p_{01} = \frac{p_{00}}{\sum_{m=0}^M OR_m \cdot p_{m0}}, \quad \text{with } OR_0 = 1. \tag{6}$$

To arrive at Equation (6), note that from the definition of the  $OR$  it follows  $OR_m p_{m0} = \frac{p_{00}}{p_{01}} \cdot p_{m1}$  for  $m = 0, \dots, M$ . Summing from  $m = 0$  to  $M$  yields  $\sum_{m=0}^M OR_m \cdot p_{m0} = \sum_{m=0}^M \frac{p_{00}}{p_{01}} \cdot p_{m1} = \frac{p_{00}}{p_{01}} \cdot \sum_{m=0}^M p_{m1} = \frac{p_{00}}{p_{01}}$  as  $\sum_{m=0}^M p_{m1} = 1$ .

For illustration, Figure 1 shows the estimated values of  $p_{01}$  for  $M = 2$ , for fixed values of  $OR_1$  and  $p_{00}$ , and varying  $OR_2$ ,  $p_{10}$ , and  $p_{20}$ . For example, with  $OR_1 = 1.2$  and  $OR_2 = 1.5$  for exposure levels 1 and 2 compared to exposure 0, population exposure frequencies of 0.3 for nonexposure, and exposure frequency increasing from 0.2 to 0.6 for exposure level 1 (corresponding to a decreasing exposure level from 0.5 to 0.1 in exposure level 2), the frequency of nonexposure in the diseased group increases from 0.23 to 0.26. Table 1 includes another example for a specific distribution (Table 1, Example 1).

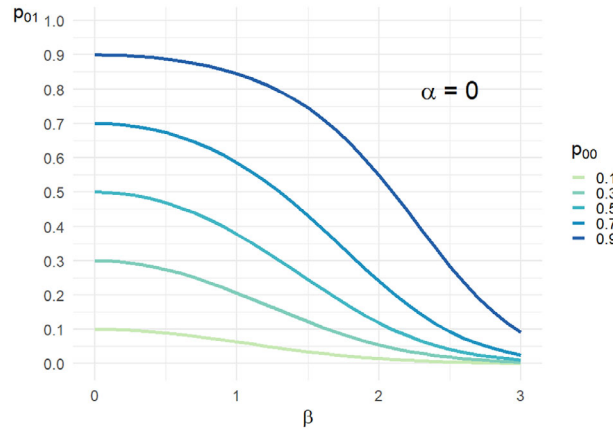


FIGURE 2 Proportion of nonexposed cases ( $p_{01}$ ) for log-normal distribution of exposure, the OR function  $\exp(\alpha + \beta \cdot \ln(x^*))$  with  $\alpha = 0$  and  $\beta$  between 0 and 3, and different proportions of nonexposed ( $p_{00}$ )

### 2.2.3 | Continuous case

The more general case is a continuous density function  $f_X(x)$  for  $X > 0$  and a continuous risk function, where we get

$$p_{01} = \frac{p_{00}}{p_{00} + \int_0^\infty OR(x) \cdot \tilde{f}_{X|Y=0}(x) dx} \tag{7}$$

Equation (7) is a direct extension of Equation (6) for the continuous case. It implies that the OR function must fulfill the condition  $\int_0^\infty OR(x) \cdot \tilde{f}_{X|Y=0}(x) dx < \infty$ . For example, if  $X$  follows a log-normal distribution, a possible OR function is  $\exp(\beta \ln(x))$ ; however,  $\exp(\beta x)$  is not valid. If  $X$  follows an exponential distribution, both  $\exp(\beta \ln(x))$  and  $\exp(\beta x)$  are possible OR functions.

Combining (3) and (7) yields

$$f_{X|Y=1}(x) = (1 - p_{00}) \frac{OR(x) \cdot f_{X|Y=0}(x)}{\int_0^\infty OR(z) \cdot \tilde{f}_{X|Y=0}(z) dz},$$

which is the density function of  $X$  in the diseased given  $X > 0$ .

For illustration, let

$$\tilde{f}_{X|Y=0}(x) = \begin{cases} p_{00} & X = 0 \\ (1 - p_{00}) \cdot \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{\ln(x)^2}{2}\right) & X > 0 \end{cases},$$

that is the positive part of  $X$  in the nondiseased is log-normally distributed with an expected value 1 and a spike at zero probability of  $p_{00}$ . Let the OR function be  $OR_{X=x^* \text{ vs } X=0} = OR(x^*) = \exp(\alpha + \beta \cdot \ln(x^*))$ . Solving Equation (7) with this density and OR function yields

$$p_{01} = \frac{p_{00}}{p_{00} + (1 - p_{00}) \cdot \exp\left(\alpha + \frac{\beta^2}{2}\right)}$$

(see Appendix). Figure 2 displays results for  $p_{01}$  for  $\alpha = 0$ , and varying values of  $\beta$  and  $p_{00}$ . For arbitrary OR functions, the integral may not be solvable, such that the solution must be obtained numerically.

Table 1 (Examples 2–7) displays results for  $p_{01}$  for some selected distributions with the expected value among the exposed in the nondiseased  $E(X|X > 0, Y = 0) = 5$  and selected OR functions. Figures A1 and A2 in the Appendix illustrate the chosen OR and density functions used in Table 1.

Since the total number of cases in a population,  $d$ , is known, the total number of nonexposed cases  $d_0$  is then given by  $d_0 = d \cdot p_{01}$ .

## 2.2.4 | Implementing estimation of $p_{01}$ and $d_0$ , including confidence intervals

R source code to calculate an estimate for  $p_{01}$  for the discrete and continuous case is available as Supporting Information on the journal's web page (<http://onlinelibrary.wiley.com/doi/10.1002/bimj.201900190/suppinfo> → link to R code Calculation\_Proportion\_Unexposed\_Cases\_p01.R). The OR function, spike probability, and conditional density function for  $X$  given  $X > 0$  have to be specified. The program is flexible and may also be used to reproduce Figure 1, Figure 2, and Table 1.

According to Equations (6) or (7), to estimate  $p_{01}$  we need an estimate of the distribution of the exposure in the population of interest, as well as estimates for the categorical ORs or the parameters of the OR function, respectively. Estimates of the population distribution may be derived from population surveys, and estimates of the ORs or the OR functions may be available from large studies or meta-analyses. For both, estimates of their standard errors are commonly available.

As  $\hat{p}_{01}$  is a complex function of the above estimates, we propose to derive a confidence interval (CI) for  $\hat{p}_{01}$  by the Monte-Carlo simulation and illustrate the procedure by two examples from Table 1.

In Example 1 of Table 1, we assumed a categorical variable  $X$  with three levels  $p_{00} = 0.3$ ,  $p_{01} = 0.5$ , and  $p_{02} = 0.2$ , and corresponding ORs of 1.5 and 2.5. Simulating 100,000 independent normally distributed parameter sets  $(\hat{p}_{00}, \hat{p}_{01}, \hat{p}_{02}, \ln(OR_1), \ln(OR_2))$  given their standard errors (0.02, 0.02, 0.02, 0.1, 0.1) and deriving the estimate of  $\hat{p}_{01}$ , we obtain a mean value for  $\hat{p}_{01}$  of 0.194 with an empirical standard error of 0.017, resulting in a 95% CI of 0.161–0.229.

In Example 3 of Table 1, we assume a log-normal distribution of  $X|X > 0$  with  $\mu = \ln(5) - 0.5$  and  $\sigma = 1$ , such that  $E(X|X > 0, Y = 0) = 5$  and for the OR function we assume  $OR(X = x \text{ vs. } X = 0) = \exp(\alpha + \beta \cdot \ln(x))$ . Let  $\hat{p}_{00} = 0.3$ ,  $\hat{\alpha} = 0.0$ ,  $\hat{\beta} = 0.5$  be the parameter estimates with standard errors of 0.02 each. Simulating again 100,000 independent normally distributed triples  $(\hat{p}_{00}, \hat{\alpha}, \hat{\beta})$  and deriving the estimate  $\hat{p}_{01}$ , we obtain a mean value for  $\hat{p}_{01}$  of 0.179 with an empirical standard error of 0.015 resulting in a 95% CI of 0.150–0.209. The estimation of  $d_0$  follows directly since  $d$  is known with  $\hat{d}_0 = d \cdot \hat{p}_{01}$ .

In practice,  $p_{01}$  is estimated separately by sex and age group, as in most cases the distribution of  $X$  is sex and age dependent. Extending the approaches is straightforward, see the data example below.

## 2.2.5 | Method II

In rare instances, a direct estimate of the incidence rate  $\lambda_0$  per year among the nonexposed is available, for example, from another population. Then trivial estimates for  $d_0$  and  $p_{01}$  are  $\hat{d}_0 = \hat{\lambda}_0 \hat{p}_{00} n$  and  $\hat{p}_{01} = \hat{d}_0 / d$ . If the rates, risk factor prevalence estimates, and population figures are given by sex  $j$  ( $j = 1, 2$ ) and age group  $k$  ( $k = 1, \dots, K$ ) as  $\lambda_{0jk}$ ,  $p_{01jk}$ , and  $n_{jk}$  (see data example), it follows that

$\hat{d}_0 = \sum_{j,k} \hat{d}_{0jk} = \sum_{j,k} \hat{\lambda}_{0jk} \cdot \hat{p}_{00jk} n_{jk}$ , and  $\hat{p}_{01jk} = \hat{d}_{0jk} / d_{jk}$ , where  $n_{jk}$  are the population figures of the nonexposed for sex  $j$  and age group  $k$  in the population of interest. The method assumes that the specific estimate of  $\lambda_0$  is available and appropriate for the population of interest. This may not be the case if there are other major risk factors for the disease, which have different distributions. This is further elaborated in the discussion and the data example.

We again propose to derive a CI for  $\hat{p}_{01}$  by the Monte-Carlo simulation. This requires standard error estimates of the prevalence of exposure as in Method I, but additionally also of all parameters of the risk function. Depending on the research setting, these might be available and then simulations can be implemented accordingly.

## 3 | EXAMPLE

In order to estimate the impact of second-hand smoke in a population, the proportion and number of never smokers among all who died of lung cancer is needed. Based on mortality statistics, the yearly number of lung cancer deaths is usually available, but there is commonly no information on the number of smokers, former smokers, and never smokers. In the following, we use the two described methods to estimate the proportion and absolute number of never smokers

among those who died of lung cancer in Germany in 2015, which is an expansion of the example presented in Becher et al. (2017). The calculations can be reproduced with the provided R script.

Smoking is a complex exposure with many characteristics influencing its effect on lung cancer risk, such as age at start of smoking, smoking dose by age, and, for former smokers, time since quitting (Gandini et al., 2008). However, public data on smoking prevalence usually do not provide such detailed information. We therefore use the classification into never ( $X = 0$ ), former ( $X = 1$ ), and current smokers ( $X = 2$ ) by sex and age group as obtained from the German Health Interview and Examination Survey for Adults (DEGS1) performed 2009–2011 (Robert Koch Institute [RKI], 2015) and sex-specific relative effect estimates for lung cancer due to current smoking  $OR_{j2}$  and former smoking  $OR_{j1}$ , relative to never smoking. Such estimates are available in the literature and show that the relative effect estimates for these smoking categories are lower in females than in males, presumably because the average smoking dose in male smokers is higher than in female smokers. For our calculations, we use the  $OR$  estimates from Gandini et al. (2008) and estimates of age- and sex-specific smoking categories from a survey in Germany (RKI, 2015) for the proportion of smokers  $p_{20,jk}$ , former smokers  $p_{10,jk}$ , and never smokers  $p_{00,jk}$  within the nondiseased. The yearly number of lung cancer deaths by sex and age is obtained from mortality statistics (RKI, 2017) and total population figures by sex and age from the German federal office of statistics (Statistisches Bundesamt, 2015).

### 3.1 | Application of Method I

According to Equation (7), the probability of being a never smoker among the lung cancer cases for sex  $j$  and age group  $k$ ,  $p_{01,jk}$  is estimated as

$$p_{01,jk} = \frac{p_{00,jk}}{p_{00,jk} + OR_{j1} \cdot p_{10,jk} + OR_{j2} \cdot p_{20,jk}}. \quad (8)$$

Given the number of lung cancer deaths in Germany by sex  $j$  and age group  $k$ ,  $d_{jk}$ , we obtain estimates for the lung cancer deaths among never smokers by sex  $j$  and age group  $k$ ,  $d_{0jk}$  as  $d_{X=0,jk} = p_{01,jk} \cdot d_{jk}$ . Based on our data, we estimate that 6,659 lung cancer deaths (95% CI 5,222–8,317) were never smokers from the total 44,813 deaths. This corresponds to a proportion of 14.9% (95% CI 11.65–18.56). Table 2 gives the age- and sex-specific estimates with CIs.

### 3.2 | Application of Method II

Several studies have provided age- and sex-specific estimates of the lung cancer mortality rate in nonsmokers ( $X = 0$ ). Combined estimates by sex (1-males; 2-females) as a continuous function of age were provided in provided in Becher et al. (2018) as updated from Winkler et al. (2011), as

$$\lambda_{01}(\text{age}) = e^{-39.8902+12.1409 \cdot \log(\text{age})-0.1155 \cdot \text{age}} / 100,000, \quad \text{and}$$

$$\lambda_{02}(\text{age}) = e^{-38.2996+11.6194 \cdot \log(\text{age})-0.1091 \cdot \text{age}} / 100,000.$$

From these, we calculated sex- and age-group specific rates based on the average age within each age interval, as German population figures and number of lung cancer deaths are given in 5-year intervals (see Appendix Table A1 for details). For comparability, we use the same age groups as in Table 2. Based on this method, we estimate that 5,294 lung cancer deaths were never smokers, which corresponds to a proportion of 11.3%. The derived age- and sex-specific estimates for the proportion of never smokers among lung cancer cases and the absolute number of never smoking lung cancer cases are presented in Table 3. As we lack measures of precision of the rate function's parameters, we could not derive corresponding CIs.

### 3.3 | Comparison of methods

A direct comparison of both methods shows that the two methods do not conclusively concur. The estimates of the total number of never smoking lung cancer deaths are higher based on Method I compared to Method II ( $\hat{d}_0 = 6,659$  vs. 5,294).

TABLE 2 Estimated proportion of never smokers in cases ( $p_{01}$ ) and lung cancer deaths among never smokers ( $d_0$ ) in Germany 2013, with simulation-based 95% CIs, by sex and age group

Sex and age group	$p_{20}$ (%) <sup>a</sup> (s.e.) smoker	$p_{10}$ (%) <sup>a</sup> (s.e.) former smoker	$p_{00}$ (%) <sup>a</sup> (s.e.) never smoker	$d_{jk}^b$	$OR_2^c$ (s.e.) of log OR	$OR_1^c$ (s.e.) of log OR	$\hat{p}_{01}$ (%) (95% CI) <sup>d</sup>	$\hat{d}_0$ (95% CI) <sup>d</sup>
Males								
<45	43.0 (1.64)	19.1 (1.35)	37.9 (1.70)	220			6.93 (5.04–9.29)	15 (11–20)
45–65	30.3 (1.79)	43.0 (1.55)	26.7 (1.43)	8,168			5.17 (3.86–6.75)	422 (315–551)
>65	11.5 (1.29)	50.8 (1.86)	37.6 (1.84)	21,296			10.00 (7.66–12.8)	2,129 (1,632–2,727)
Total				<b>29,684</b>	<b>9.87 (0.187)</b>	<b>4.43 (0.167)</b>	<b>8.65 (6.59–11.11)</b>	<b>2,566 (1,958–3,297)</b>
Females								
<45	34.9 (2.27)	17.9 (1.12)	47.2 (1.71)	139			12.70 (9.46–16.58)	18 (13–23)
45–65	27.8 (1.60)	30.3 (1.33)	41.9 (1.53)	4,823			11.83 (9.02–15.10)	571 (435–728)
>65	9.0 (1.09)	20.0 (1.79)	71.1 (1.99)	10,167			34.46 (27.70–41.99)	3,504 (2,816–4,269)
Total				<b>15,129</b>	<b>7.58 (0.177)</b>	<b>3.35 (0.167)</b>	<b>27.05 (21.58–33.18)</b>	<b>4,093 (3,264–5,020)</b>
Total				44,813			<b>14.86 (11.65–18.56)</b>	<b>6,659 (5,222–8,317)</b>

Abbreviation: CI, confidence interval; s.e., standard error.

<sup>a</sup>Robert Koch Institute (2015).

<sup>b</sup>Robert Koch Institute (2017).

<sup>c</sup>Gandini et al. (2008).

<sup>d</sup>Monte-Carlo simulation.



**TABLE 3** Estimated proportion of never smokers in cases ( $p_{01}$ ) and absolute number of lung cancer cases by smoking status, based on nonsmoker rates (Method II) and observed numbers, Germany, 2013

Sex	Age group	$\hat{p}_{01}$ (%)	Never smoker <sup>a</sup> $\hat{d}_0$	Former smoker <sup>a</sup> $\hat{d}_1$	Current smoker <sup>a</sup> $\hat{d}_2$	Total <sup>a</sup> $\hat{d}$	$d_{jk}$ <sup>b</sup>
Males	<45	14.09	31	68	340	439	220
	45–65	4.77	382	2,731	4,286	7,399	8,168
	>65	6.37	1,357	8,123	4,095	13,575	21,296
	Total	<b>5.96</b>	<b>1,770</b>	<b>10,922</b>	<b>8,721</b>	<b>21,413</b>	<b>29,684</b>
Females	<45	25.18	35	44	192	271	139
	45–65	10.86	524	1,270	2,637	4,431	4,823
	>65	29.16	2,965	2,793	2,844	8,602	10,167
	Total	<b>23.29</b>	<b>3,524</b>	<b>4,107</b>	<b>5,673</b>	<b>13,304</b>	<b>15,129</b>
Total		11.81	5,294	15,029	14,394	34,717	44,813

<sup>a</sup>Derived based on prevalence of smoking status as in RKI (2015), the incidence rate function by age proxies as provided by Becher et al. (2018) as modified from Winkler et al. (2011) and population figures as provided in Statistisches Bundesamt (2015).

<sup>b</sup>Statistisches Bundesamt (2015).

Over all age groups, our results suggest that the proportion of never smokers among the lung cancer cases is higher among women than among men, which we think is reasonable due to the higher prevalence of never smoking women. Only within the younger age group, the estimated proportion of never smokers is higher based on Method II and therefore the absolute number somewhat lower—for both males and females (Figure 3). Here Method II results in a much higher proportion of cases than in the middle age 45–65, which seems implausible. In our example, this is probably due to an overestimation of the lung cancer rate in the young age group. On the other hand, the rates are generally low at young ages and contribute little to the total number.

All of this, however, does not indicate which of the two estimates is closer to the true value. To compare the methods in more detail, we can additionally compare the number of former and current smokers among the lung cancer cases as derived by the two methods, where we see the same trend in lower estimates for the two higher age groups based on Method II (Appendix Figure A3).

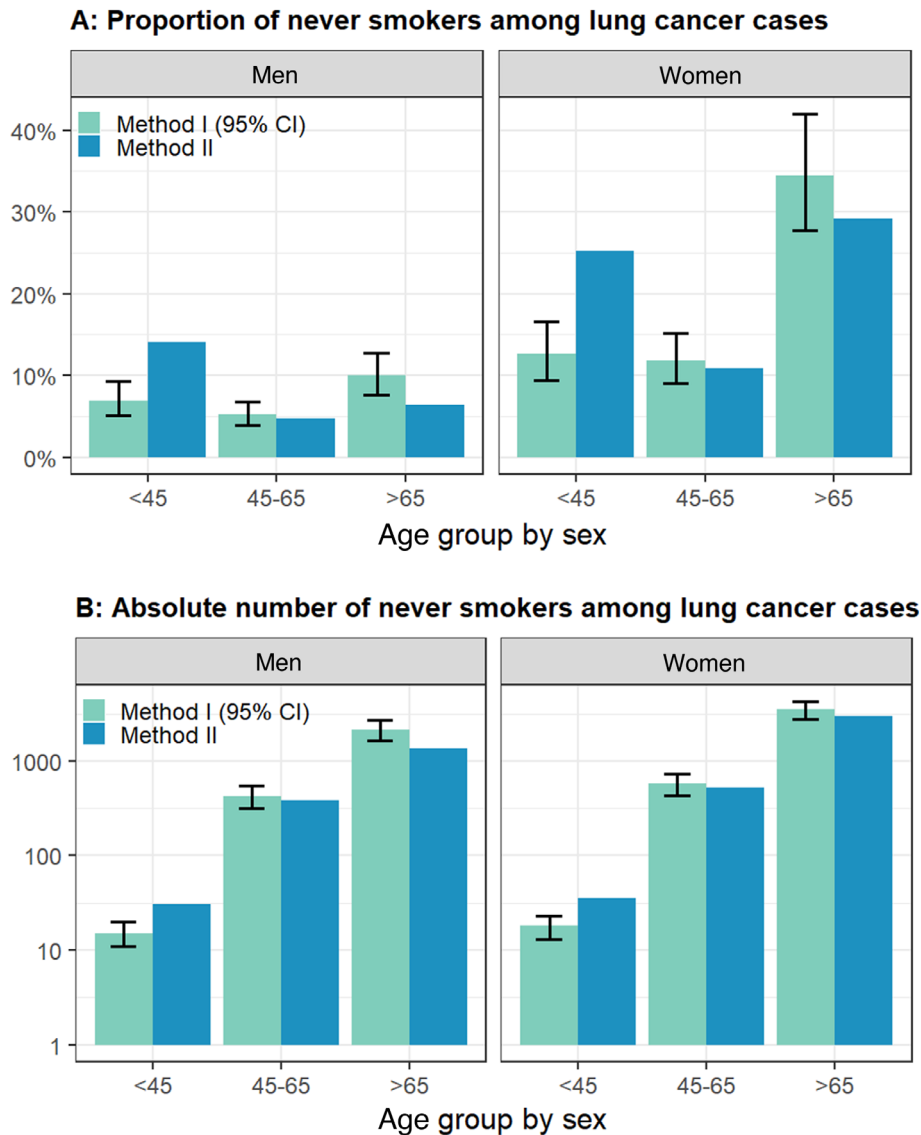
R source code is available as Supporting Information on the journal's web page (<http://onlinelibrary.wiley.com/doi/10.1002/bimj.201900190/suppinfo>) which covers the full data example of second-hand smoking, including the data and can be used to reproduce Figure 3.

## 4 | DISCUSSION

We presented two methods to estimate the number of nonexposed cases in a population. Methods I and II both require knowledge of the risk factor distribution in the nondiseased, as well as population figures for the nondiseased. The difference between the two methods is that Method I additionally requires an estimate of the risk function for a given exposure relative to the nonexposed, whereas Method II requires an estimate of the incidence/mortality rate in the nondiseased. An estimate for the rate may generally not be available, that is Method II has practical limitations. For the presented example of lung cancer deaths in Germany, with smoking as the exposure of interest, this estimate was available and—as such—motivated the present paper. However, the estimates needed for Method I are available more frequently. For this method, we developed an R script to calculate the estimated number of nonexposed cases.

Currently, population and mortality registries in Germany generally do not include information on previous exposures other than sex and age, and we are not aware of any other national registries which do so. A change in public health policy resulting in this information being included for mortality registries would render the presented methods superfluous. However, such a policy change will take a substantial amount of time, and depending on the way this data was collected, the reliability of it needs to be checked before it can be confidently used for research and other purposes.

In general, the estimates resulting from these methods are conceptually different, and are only unbiased if the underlying assumptions are fulfilled. The assumptions underlying any of these procedures must be considered carefully. Method I implies that the sum of the estimates of exposed and nonexposed cases yields exactly the known total number of cases. In other words, if the total number of nonexposed cases is underestimated (e.g., if the relative effect is overestimated), the total number of exposed cases is overestimated, and vice versa. Method II, on the other hand, also allows estimation of



**FIGURE 3** Comparison of Methods I and II to estimate the proportion of never smokers among lung cancer cases, just as the absolute number of never smoking lung cancer cases in Germany, 2013

the number of exposed cases, if an effect estimate of the exposure is available. It is then possible to compare the sum of the estimated number of cases over all exposure categories. In our example, the largest difference is found for the old age group in males (13,575 vs. 21,296). This may indicate that the disease rate estimate underestimates the true risk in older males; however, no data are available to further check this.

Furthermore, the incidence or mortality rate of a certain disease usually depends on multiple factors. For chronic diseases, latency periods play a major role, and the prevalence of an exposure may have an impact on the mortality due to a certain disease only after some time. The estimate of a disease's baseline rate among the nonexposed may be obtained from a different population, for example, with a different distribution of other risk factors. Although conceptually appealing, this seems to be a substantial limitation of Method II. The estimates needed for Method I, on the other hand, seem less critical. Relative effect estimates, especially if obtained from large studies or meta-analyses, may be accepted as relatively precise.

Prevalence estimates for the risk factor of interest, as needed for both methods, may be available from surveys. However, two points need to be considered here. First, smoking is a multidimensional and continuous risk factor, and therefore a major limitation lies in the fact that survey data, upon which prevalence estimates are based, usually only work with broad categories, for example, grouping individuals into never, former, and current smokers. By estimating relative effects based on these categories, there is a subsequent correspondence to the average dose within a group. If, for example, the mean cumulative dose is higher in the older age groups, a higher relative effect should be used. This would increase the estimated

total number of cases in Method I. We think, however, that the risk estimates are relatively precisely known, especially compared to the other two parameters, the age- and sex-specific smoking prevalence estimates and the lung cancer rate in never smokers. Second, the latency period from exposure to disease onset may cause considerable complications. This is particularly relevant if the distribution changes over time, as is the case for smoking in many populations. In the example and in the previous paper, we simply assumed that the deaths in 2013 can be modeled using the prevalence estimates from 2009–2011 which is quite a simplification.

In our example, the prevalence estimates were taken from a national survey and the *OR* estimates from a meta-analysis. For both the approximate standard errors but no original data were available, such that we employed a Monte-Carlo simulation to derive a CI of our parameter of interest based on these standard errors. If in another setting original data are available, we would recommend to employ a bootstrap approach to derive a CI instead.

We presented two estimation methods for possible application in other settings. As a general recommendation regarding the choice of a method is not possible, the decision regarding method must be adapted to the specific situation at hand. If estimates for all parameters are available, both methods should be used and compared against each another. Additionally, issues regarding the reliability of individual estimates should be considered prior to the interpretation of results.


## ACKNOWLEDGMENTS

Open access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Heiko Becher  <https://orcid.org/0000-0002-8808-6667>

## REFERENCES

- Becher, H., Belau, M., Winkler, V., & Aigner, A. (2018). Estimating lung cancer mortality attributable to second hand smoke exposure in Germany. *International Journal of Public Health, 63*(3), 367–375.
- Becher, H., Lorenz, E., Royston, P., & Sauerbrei, W. (2012). Analysing covariates with spike at zero: A modified FP procedure and conceptual issues. *Biometrical Journal, 54*, 686–700.
- Federal Statistical Office (2019). Retrieved from [https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Todesursachen/\\_inhalt.html](https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Todesursachen/_inhalt.html)
- Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., & Boyle, P. (2008). Tobacco smoking and cancer: A meta-analysis. *International Journal Cancer, 122*, 155–164.
- Heuschmann, P. U., Heidrich, J., Wellmann, J., Kraywinkel, K., & Keil, U. (2007). Stroke mortality and morbidity attributable to passive smoking in Germany. *European Journal of Cardiovascular Prevention & Rehabilitation, 14*(6), 793–795.
- Mittleman, M. A. (1995). Estimation of exposure prevalence in a population at risk using data from cases and an external estimate of the relative risk. *Epidemiology, 6*(5), 551–553.
- Robert Koch-Institut (Hrsg) (2017). *Krebs in Deutschland für 2013/2014. II*. Berlin, Germany: Ausgabe. [https://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs\\_in\\_Deutschland/kid\\_2017/krebs\\_in\\_deutschland\\_2017.pdf?\\_\\_blob=publicationFile](https://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/kid_2017/krebs_in_deutschland_2017.pdf?__blob=publicationFile)
- Robert Koch-Institute (RKI), Department of Epidemiology and Health Monitoring. (2015). *German Health Interview and Examination Survey for Adults (DEGSI)*. Public Use File first Version. <https://doi.org/10.7797/16-200812-1-1-1>
- Royston, P., Sauerbrei, W., & Becher, H. (2010). Modelling continuous exposures with a “spike” at zero: A new procedure based on fractional polynomials. *Statistics in Medicine, 29*, 1219–1227.
- Schelchase, T., & Weber, S. (2007). Mortality statistics in Germany. Problems and perspectives. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz, 50*(7), 969–976.
- Sitas, F., Bradshaw, D., Egger, S., Jiang, G., & Peto, R. (2018). Smoking counts: Experience of implementing questions on smoking on official death certification systems. *International Journal of Epidemiology, 48*(2), 633–639.
- Statistisches Bundesamt (2015) *Statistisches Jahrbuch 2015 für die Bundesrepublik Deutschland*. Wiesbaden, Germany: Author.
- Winkler, V., Ng, N., Tesfaye, F., & Becher, H. (2011). Predicting lung cancer deaths from smoking prevalence data. *Lung Cancer, 74*, 170–177.
- World Health Organization. (2019). Mortality database updated. Retrieved from <http://www.euro.who.int/en/data-and-evidence/archive/mortality-database-updated>.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Becher H, Aigner A. Methods to estimate proportion and number of nonexposed cases in a population. *Biometrical Journal*. 2021;63:514–527. <https://doi.org/10.1002/bimj.201900190>

**APPENDIX**

According to Equation (7)  $p_{01} = \frac{p_{00}}{p_{00} + \int_0^\infty OR(x) \tilde{f}_{X|D=0}(x) dx}$ , we have to solve  $\int_0^\infty OR(x) \tilde{f}_{X|D=0}(x) dx$

With  $OR(x) = \exp(\alpha + \beta \ln(x))$  and  $f_{X|D=0}(x) = \frac{1}{\sqrt{2\pi x}} \exp(-\frac{\ln(x)^2}{2})$ , we get

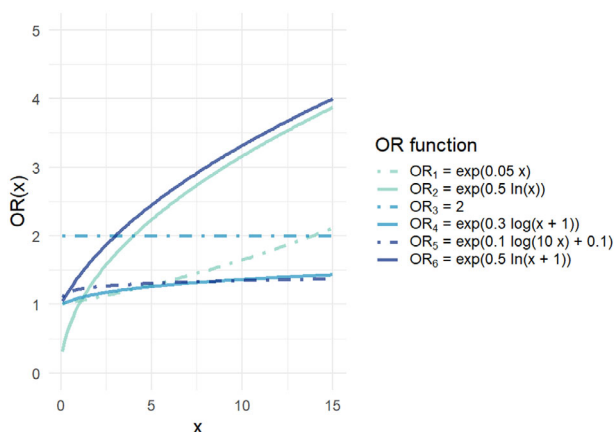


FIGURE A1 Plot of OR functions as used in Table 1

TABLE A1 Method II: Estimated lung cancer deaths among never smokers ( $d_{0jk}$ ) in Germany 2013, by sex  $j$  and age group  $k$ , based on population figures  $n_{jk}$ , estimated lung cancer rates  $\lambda_{jk}$  (rates per 100,000 per year), and number of lung cancer deaths ( $d_{jk}$ )

Age group	Men				Women			
	$n_{1k}$	$\lambda_{01k}$ per 100,000 <sup>a</sup>	$d_{1k}$ <sup>b</sup>	$\hat{d}_{01k}$ <sup>c</sup>	$n_{2k}$	$\lambda_{02k}$ per 100,000 <sup>a</sup>	$\hat{d}_{02k}$ <sup>b</sup>	$d_{02k}$ <sup>c</sup>
<25	2,290,000	0.01	4	0	2,171,000	0.01	5	0
25–29	2,720,000	0.06	9	1	2,567,000	0.06	4	1
30–34	2,604,000	0.25	16	2	2,497,000	0.25	8	3
35–39	2,481,000	0.80	39	8	2,425,000	0.76	28	9
40–44	2,504,000	2.06	152	20	2,457,000	1.88	94	22
45–49	3,291,000	4.47	558	39	3,211,000	3.97	412	53
50–54	3,508,000	8.45	1,385	79	3,440,000	7.35	960	106
55–59	3,012,000	14.31	2,541	115	3,024,000	12.26	1,504	155
60–64	2,530,000	22.10	3,684	149	2,673,000	18.72	1,947	210
65–69	2,082,000	31.58	4,017	247	2,253,000	26.54	2,069	425
70–74	1,848,000	42.21	6,077	293	2,125,000	35.28	2,522	533
75–79	1,882,000	53.24	5,265	377	2,376,000	44.38	2,131	750
80–84	1,015,000	63.84	3,747	244	1,502,000	53.18	1,696	568
>85	647,000	80.68	2,190	196	1,435,000	67.50	1,749	689

<sup>a</sup>Based on formula by Becher et al. (2018) as modified from Winkler et al. (2011).

<sup>b</sup>Statistisches Bundesamt (2015).

<sup>c</sup>Additionally, based on prevalence estimates by age and sex group as in Table 1.

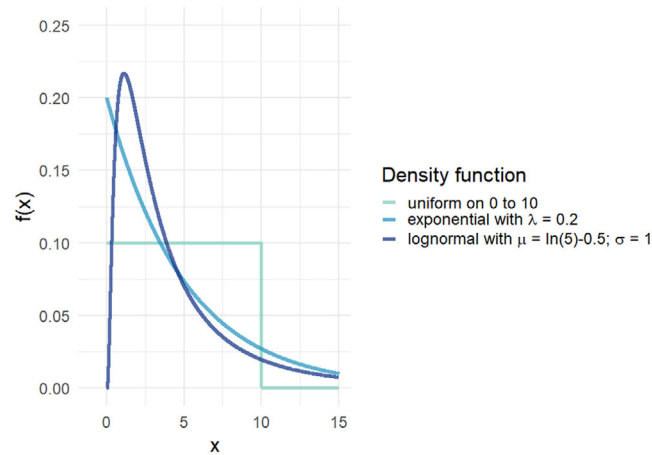


FIGURE A2 Plot of density functions as used in Table 1

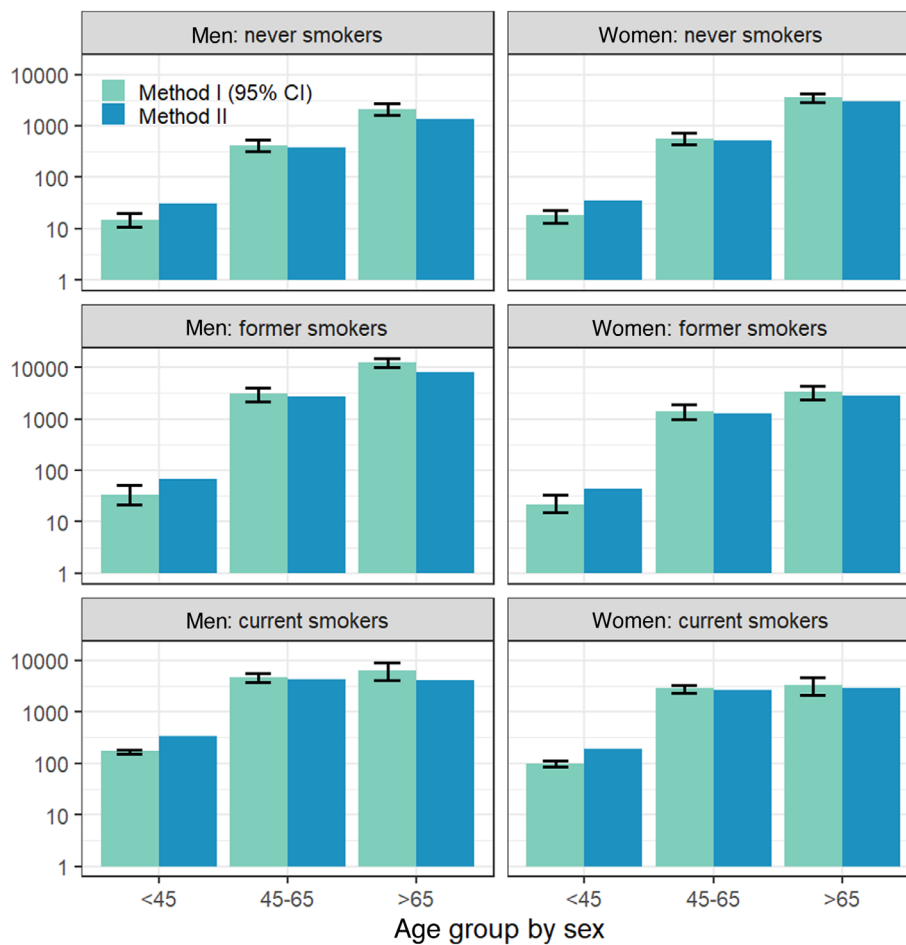


FIGURE A3 Comparison of Methods I and II to estimate the absolute number of never, former and current smokers among lung cancer cases in Germany, 2013

$\tilde{f}_{X|D=0}(x) = (1 - p_{00}) \frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{\ln(x)^2}{2}\right)$  and it follows:

$$\int_0^{\infty} OR(x) \tilde{f}_{X|D=0}(x) dx = \int_0^{\infty} \exp(\alpha + \beta \ln(x)) (1 - p_{00}) \frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{\ln(x)^2}{2}\right) dx$$

$$\begin{aligned}
&= (1 - p_{00}) \exp(\alpha) \int_0^{\infty} \frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{\ln(x)^2}{2} + \beta \ln(x)\right) dx \\
&= (1 - p_{00}) \exp(\alpha) \int_0^{\infty} \frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{(\ln(x) - \beta)^2}{2} + \frac{\beta^2}{2}\right) dx \\
&= (1 - p_{00}) \exp\left(\alpha + \frac{\beta^2}{2}\right) \int_0^{\infty} \frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{(\ln(x) - \beta)^2}{2}\right) dx \\
&= (1 - p_{00}) \exp\left(\alpha + \frac{\beta^2}{2}\right)
\end{aligned}$$

Inserting the result in Equation (7) yields

$$\begin{aligned}
p_{01} &= \frac{p_{00}}{p_{00} + (1 - p_{00}) \cdot \exp\left(\alpha + \frac{\beta^2}{2}\right)} \\
&= \frac{1}{1 + \frac{(1-p_{00})}{p_{00}} \cdot \exp\left(\alpha + \frac{\beta^2}{2}\right)}.
\end{aligned}$$