Anhang 5

A methodology to assess possible effects of enhanced surveillance on the risk estimate from ecologic studies of thyroid cancer after the Chernobyl accident

3 December 2004

A methodology to assess possible effects of enhanced surveillance on the risk estimate from ecologic studies of thyroid cancer after the Chernobyl accident

J. Christian Kaiser*,[1], Peter Jacob[1] and Sergey Vavilov[2]

[1] GSF National Research Center, Institute of Radiation Protection,
D-85764 Neuherberg, Germany
[2] Radiation Protection Institute, Academy of Technological Sciences,
252050 Kyiv, Ukraine

Summary

For two reasons quantitative post-Chernobyl assessment of the thyroid cancer risk from ^{131}I exposure in children under age 18 at the time of accident has been carried out with spatially aggregated data. Firstly, individual data has not been available up to now. Secondly, a large number of individuals can be included in the analysis for a better statistical power. But aggregation can destroy important individual information. Therefore, the risk estimate from an ecologic study may be biased, especially in the presence of confounding factors. In 1998 Lubin has investigated such an ecologic bias for lung cancer caused by radon exposure confounded by smoking. We generalise his approach for the Chernobyl case where enhanced medical surveillance (or screening) has been identified as the most important confounder of cancer incidence data. Our investigations were performed with both Monte-Carlo simulations and analytical calculations, respectively. Based on realistic dose estimates for 743 Ukrainian settlements we simulated individual data sets on exposure, screening and health status for 366397 children. To generate the cancer cases, an individual risk model with linear dose response has been combined with three different screening models. Poisson regression on a linear model for a mean settlement risk produced an ecologic bias in most cases, when screening information was neglected. The estimated risk parameters deviate from the true mean parameters which describe the true risk in the population. The bias is caused by those correlations between screening and risk factors, which have not been included in the definition of the true population-based risk. Analytical relations have been derived to calculate the bias numerically exact from the population data.

Key words: Risk estimation; Ecologic bias; Screening effect; Thyroid cancer; Chernobyl accident

* Corresponding author: e-mail: `christian.kaiser@gsf.de`

1. Introduction

Recently, a number of studies (Buglova et al. 1996; Jacob et al. 1998; Jacob et al. 1999; Likhtarev et al. 1999) on the thyroid cancer risk of children exposed to ^{131}I after the Chernobyl accident has been performed for contaminated areas of the former Soviet Union. They were based on aggregate exposure data to increase the number of persons at risk for a better statistical power. Astakova et al. (1998) have carried out a case-control study in Belarus for a limited number of individuals. But sufficient individual data for the total study population is not (and will probably never be) available. The results from the ecologic studies supported a linear dose-response relationship and gave estimates of the excess absolute risk per unit thyroid dose (EARPD). However, Ron et al. (1992) have emphasised that a good control of possible confounding factors is essential to derive reliable risk estimates. Enhanced thyroid surveillance (or screening) has been identified as the most important factor, but due to a lack of information a thorough quantitative assessment was not feasible. This was the motivation to derive here a methodology for the assessment of possible screening effects on the risk estimate.

After the Chernobyl accident Belarus and Russia decided to set up yearly medical surveillance programmes for exposed children. Thus, thyroid cancer incidence may be enhanced by screening in the form of more frequent medical check-ups, refined examination methods like thyroid hormone testing or ultrasound imaging. Case reporting to the central registries has been improved and an enhanced public risk awareness has also contributed to a higher cancer detection rate. If in a cohort study all members have been examined with the same examination method screening effects are possible but they are not correlated to radiation exposure. But in Belarus mobile medical teams have checked children more thoroughly in highly contaminated regions. In this case a screening factor would be positively correlated to the thyroid dose. Ongoing research in the Ukraine revealed significant negative dose-screening correlations on an oblast level.

A wealth of information on the limitations of risk estimates from ecologic studies is available in the epidemiologic literature (Morgenstern 1998; Greenland 2001; Greenland 2002). For long it is known that an ecologic bias of cross-level inference may arise if risk parameters, which have been produced with aggregate data, are assigned to individuals (Firebaugh 1978). Parameter estimates for linear risk models may depend on the group design even for un-confounded relations between predictor and outcome variables (Piantadosi et al. 1988). The NCRP (2001) report no. 136 lists several potential weaknesses of ecologic studies like inadequate use of summary variables as confounders or no control of misclassification.

If a predicting variable like exposure is accompanied by a confounding risk factor, Piantadosi (1994) noticed that the covariance between these two variables controls the ecologic bias. This bias has been further quantified for the problem of lung cancer risk related to radon exposure confounded by smoking (Lubin 1998; Lubin 2002). The mathematical structure of this problem is similar to that of thyroid cancer risk from iodine exposure for children after Chernobyl confounded by screening. There is, however, a conceptual difference because lung cancer can be caused biologically by both radiation exposure and smoking. On the other hand screening does not induce cancer but increases the number of thyroid cancer cases that are reported to the registries.

To assess the screening effect we simulated individual data sets on exposure, screening and health status for 366397 Ukrainian children. The individual thyroid doses were simulated based on preliminary dose estimates for 743 Ukrainian settlements. Approximate values for the screening factor are given in Jacob et al. (1999) and the references therein. The cancer cases were generated from an individual risk model with constant background and a linear dose response that has been combined with three different screening models.

Poisson regression on the number of simulated cancer cases to a settlement-based (or ecologic) linear risk model produced a bias in most cases when screening information has been neglected. We consider the risk parameters of the ecologic model as biased if they deviate from the true mean risk parameters which could be assigned to the study population if all individual information on screening and exposure were available. In the Appendix analytical relations have been derived to calculate this bias numerically exact.

2. Mean settlement risk

Let n_c be the total number of thyroid cancer cases found in a study area with a population of N_p individuals at risk during an observation time $\Delta T = 10$ yr from 1990-99. We assume that this number can be decomposed into four components

$$n_c = n_{0n} + n_{0s} + n_{rn} + n_{rs}. \tag{1}$$

The n_{0n} spontaneous cases would have been found during the observation time in a situation without the accident and without enhanced surveillance. The n_{rn} radiation-induced cases would have become clinically relevant after the accident if the surveillance regime had been left unchanged. The $n_{0s} + n_{rs}$ additional cases may be attributed to the extended medical check-up programmes after 1986.

Motivated by the decomposition of the total number of cases in Equation (1), the risk of contracting a thyroid tumour is expressed as

$$h_{ij} = (1 + \eta_{ij})h_0 + (1 + \kappa_{ij})\beta D_{ij} \tag{2}$$

for an individual j in a geographical unit i. From here on this unit is called a settlement with N_i persons at risk. The constant risk factor β would describe the risk in the population after the accident without any change in the medical surveillance regime. For simplification a constant background risk h_0 is applied to the whole study population and possible age and/or sex dependencies are neglected. The effect of screening influences the recorded incidence of spontaneous and radiation-induced thyroid tumours in a different way if, for example, a large number of occult spontaneous cancer cases were detected. Therefore, two individual screening factors η_{ij} and κ_{ij} enhance either the spontaneous or the radiation-induced risk. These factors were zero if the medical surveillance regime had been left as it was before the accident.

It is impossible to collect individual data on screening and exposure for more than 300000 exposed children in the Ukraine. However, mean dose values for settlements will be available soon (see Subsection 4.1). Therefore, a treatment of the problem on a settlement level looks more promising. The corresponding mean risk for a settlement *i* becomes

$$\bar{h}_i = (1+\bar{\eta}_i)\,h_0 + (1+\bar{\kappa}_i)\,\beta\bar{D}_i\left(1+\frac{\mathrm{cov}_{Ii}(\kappa,D)}{(1+\bar{\kappa}_i)\,\bar{D}_i}\right). \tag{3}$$

Equation (3) introduces the *intra* settlement covariance $\mathrm{cov}_{Ii}(\kappa,D)$ of the radiation screening factor and the dose. It is derived in Appendix A. Lubin (1998) pointed out that neglecting $\mathrm{cov}_{Ii}(\kappa,D)$ in the regression causes a bias from ill-conducted cross-level inference.

The observed cancer cases were regressed on the mean settlement dose $\bar{D}_i$ with the mean settlement risk

$$\bar{h}_{i,eco} = h_{0,eco} + \beta_{eco}\bar{D}_i. \tag{4}$$

By using $\bar{h}_{i,eco}$ instead of $\bar{h}_i$ in the regression an additional bias enters the risk analysis, because Equation (4) ignores any information on confounders like screening. According to Stidley and Samet (1994) this bias arises from a model misspecification. Equation (4) constitutes the risk model that introduces the ecologic bias in the present work.

3. Ecologic bias

If all individual information on screening and exposure were available, the risk parameter

$$\beta_{pop} \equiv (1+\langle\kappa\rangle)\beta\left(1+\frac{\langle\mathrm{cov}_I(\kappa,D)\rangle + \mathrm{cov}_S(\bar{\kappa},\bar{D})}{(1+\langle\kappa\rangle)\,\langle D\rangle}\right) \tag{5}$$

would describe the true mean EARPD in the study population. A derivation is given in Appendix A also for the mean background risk $\langle h_0\rangle_{pop}$.

The aim of an ecologic study is to produce an estimate for β_{pop} (and $\langle h_0\rangle_{pop}$) to quantify the number of excess thyroid cancer cases after the accident in the exposed population. In the event of another accident they will help to predict the number of expected excess cases elsewhere. This will facilitate the planning of remediation measures to provide adequate medical treatment.

If all individuals were selected at random from the population, no correlations between exposure to ^{131}I and screening would occur. For this special case (Appendix B.2) an ecologic regression with Equation (4) will yield β_{eco} as the correct estimate for β_{pop}.

In general, one cannot exclude possible correlations of screening, exposure and background risk either within settlements or between settlements. Among others, they produce the covariances $\langle\mathrm{cov}_I\rangle$ (A.4) and cov_S (A.9), respectively. Now the ecologic regression will give an EARPD $\beta_{eco} \neq \beta_{pop}$ in most cases.

The EARPD β_{pop} would quantify the true population-based risk under the very unlikely condition, that all individual information on screening and exposure were available. The EARPD β_{eco} from ecologic regression may deviate from β_{pop} and therefore cause the ecologic bias in the post-Chernobyl risk assessment of thyroid cancer.

Lubin (1998) used a different definition of the ecologic bias which he attributed to the cross-level inference from the individual to the aggregate level. In our case both β_{eco} and β_{pop}

are defined on the level of the total study population. Nevertheless, in both applications the bias is caused by correlations between the risk factors and a confounding variable.

4. Materials and Methods

To simulate population data for radiation exposure, enhanced medical surveillance and health status three stages are necessary as shown in Figure 1. They are described in turn in the next three subsections.

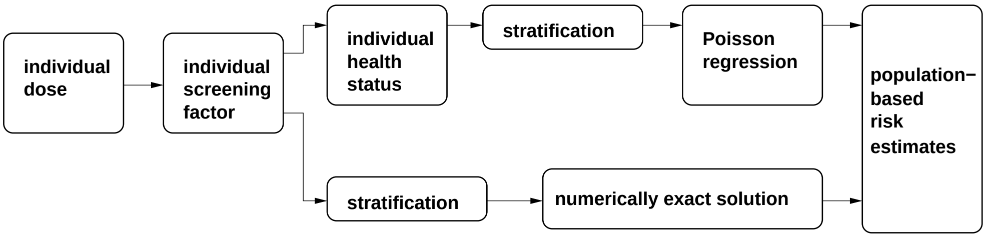


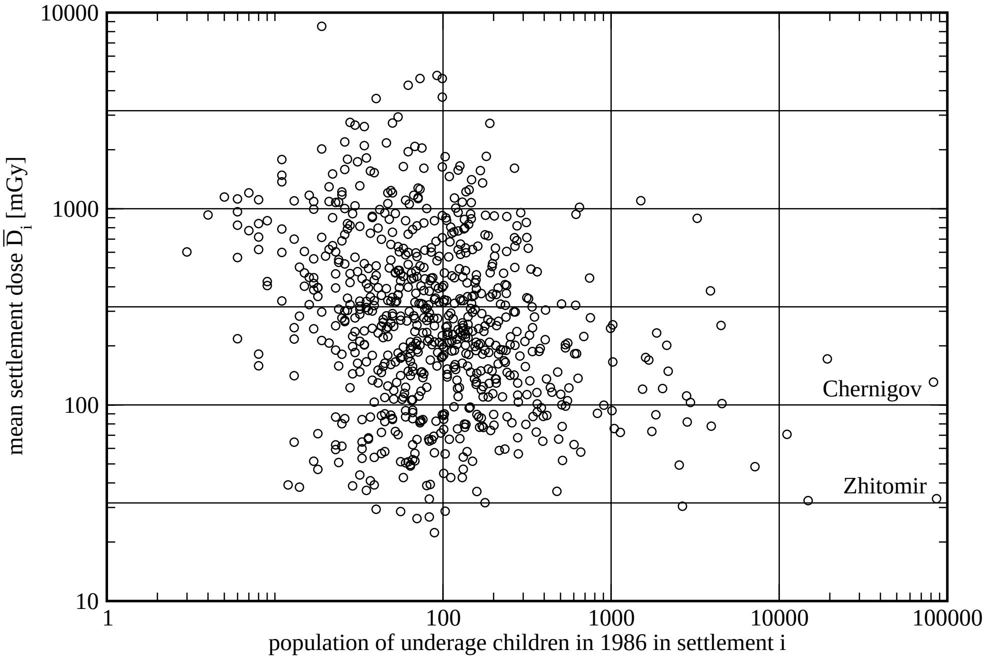
Figure 1. Simulation scheme for one population data set, risk estimates are obtained (after stratification) either with Poisson regression on the number of simulated cases or from a numerically exact solution; to gain accuracy for the risk estimates this scheme is repeated 100 times

4.1. Individual dose simulation

Individual thyroid doses have been simulated for 366397 Ukrainian children in 18 birth year groups from 1968-85 whose place of residence was known for 743 settlements in 1986 (USSR 1991). The sizes ranged from the two largest cities of Zhitomir and Chernigov with more than 80000 children to small villages with below 100 children. The study population comes mostly from rural areas without the city of Kiev. For simplification we did not distinguish between sexes. For each birth year group of a settlement we assumed a lognormal dose distribution with a geometric standard deviation (GSD) of 2.2. The geometric means are the unpublished preliminary estimates of the Ukrainian Radiation Protection Institute which are produced in an ongoing research project together with the GSF - Institute of Radiation Protection. They are always decreasing with age due to the higher mass of the thyroid gland. Thus, the individual doses were drawn from 18 lognormal thyroid dose distributions in each settlement. The 18 dose means for the birth year groups can be aggregated to a settlement dose mean $\overline{D}_i$. The 743 arithmetic settlement dose means are shown in Figure 2 (top). The mean dose per person and the measured maximal dose were approx. 173 mGy and 10000 mGy, respectively. Figure 2 (bottom) depicts a typical simulated cumulative dose distribution.

4.2. Screening models with correlation to exposure

Selecting realistic models for the correlation between screening and exposure is a difficult task. Thus, our models have been invented mainly for didactic reasons. For simplification we



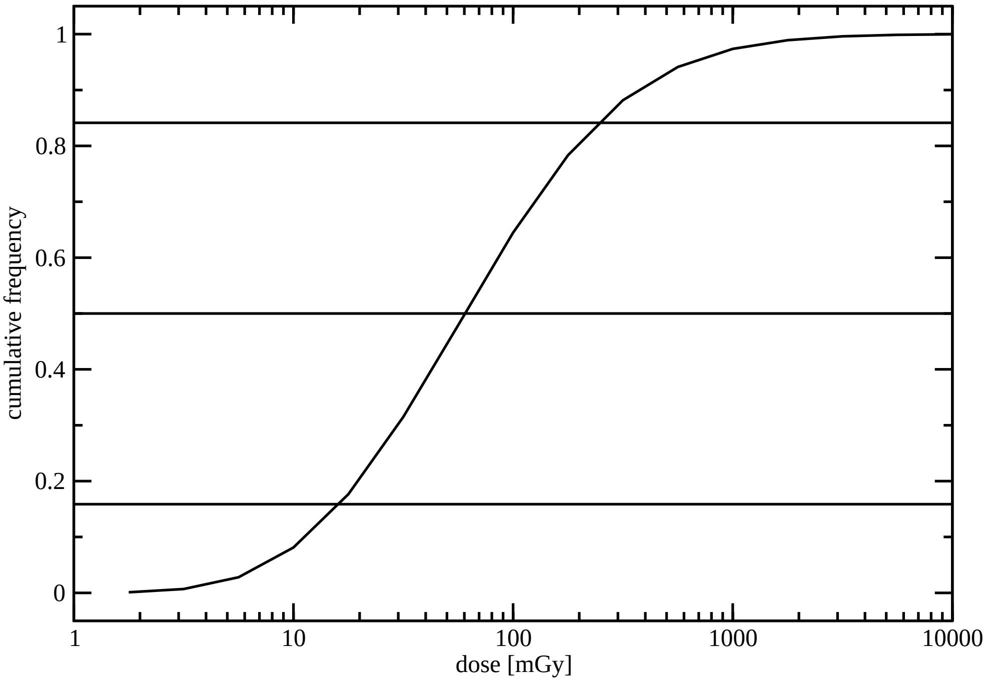


Figure 2. Top: dose means for 743 Ukrainian settlements; bottom: simulated cumulative dose distribution for 366397 individuals with arithmetic mean 173 mGy and arithmetic standard deviation 515 mGy

chose the same individual screening factor

$$\eta_{ij} = \kappa_{ij} \tag{6}$$

for the spontaneous and the radiation dependent part of the individual risk from Equation (2). It should be noted, however, that in the Chicago hospital study of Ron et al. (1992) $\langle\eta\rangle$ was larger than $\langle\kappa\rangle$.

Screening above an individual dose level: individual model. If the thyroid doses for all children were known one could fix a population-wide level D_{scrn} for the individual dose. Children with a dose $D_{ij} > D_{scrn}$ would enter the medical check-up programme. The proportion of screened children can then be taken from Figure 2 (bottom). This model is called the individual (screening) model. It produces correlations within a settlement and between settlements, i.e. cov_{Ii}, $\text{cov}_S \neq 0$.

Screening above a settlement dose level: settlement model. In a more realistic scenario only mean doses for settlements were known (Figure 2, top). Now D_{scrn} denotes the settlement dose level and *all* children from a settlement with $\bar{D}_i > D_{scrn}$ would be screened, irrespectively of their individual doses. For this settlement (screening) model correlations only appear between settlements but not within a settlement, i.e. $\text{cov}_{Ii} = 0$, $\text{cov}_S \neq 0$.

Screening of a fixed proportion from a settlement: percentile model. If in each settlement the same percentile z_p of children with the highest doses are screened, only intra settlement correlations arise, i.e. $\text{cov}_{Ii} \neq 0$. This percentile (screening) model seems contrived but it provides valuable insight because there is no inter settlement correlation, i.e. $\text{cov}_S = 0$. Lubin (2002) used a similar model of proportions of exposed individuals to describe smoking information.

For an individual j in a settlement i we set

$$\kappa_{ij} = \kappa_{ind}\,\theta_{ij} \quad \text{with} \quad \theta_{ij} = \begin{cases} 1 & : \quad \text{if it is screened} \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{7}$$

The factor κ_{ind} is used to scale the screening strength and has been set here mostly to one or two. For the Chernobyl case Jacob et al. (1999) have estimated population-wide values $\langle\kappa\rangle = \kappa_{ind} z_p$ of 0 to 4 in Belarus. Their factor S equals $1 + \langle\kappa\rangle$ in this paper. Ron et al. (1992) report a maximal factor of 11 for a cohort of patients in a screening programme at a Chicago hospital.

Figure 3 shows simulated covariances for the three screening models as a function of the proportion (or percentile) of screened individuals z_p. For the individual and the settlement screening model this proportion is determined by the corresponding dose level D_{scrn}. Specific values are given in Tables C1, C2 and C3 in Appendix C. All covariance curves show a similar behaviour. If for high dose levels the proportion z_p is small, the covariances are also small because of the low number of screened persons. The curves reach their maxima for z_p between 0.1 and 0.5 and tend to zero if all individuals are screened. The Equations (A.4) and (A.9) show that the covariances are proportional to the individual screening factor κ_{ind}.

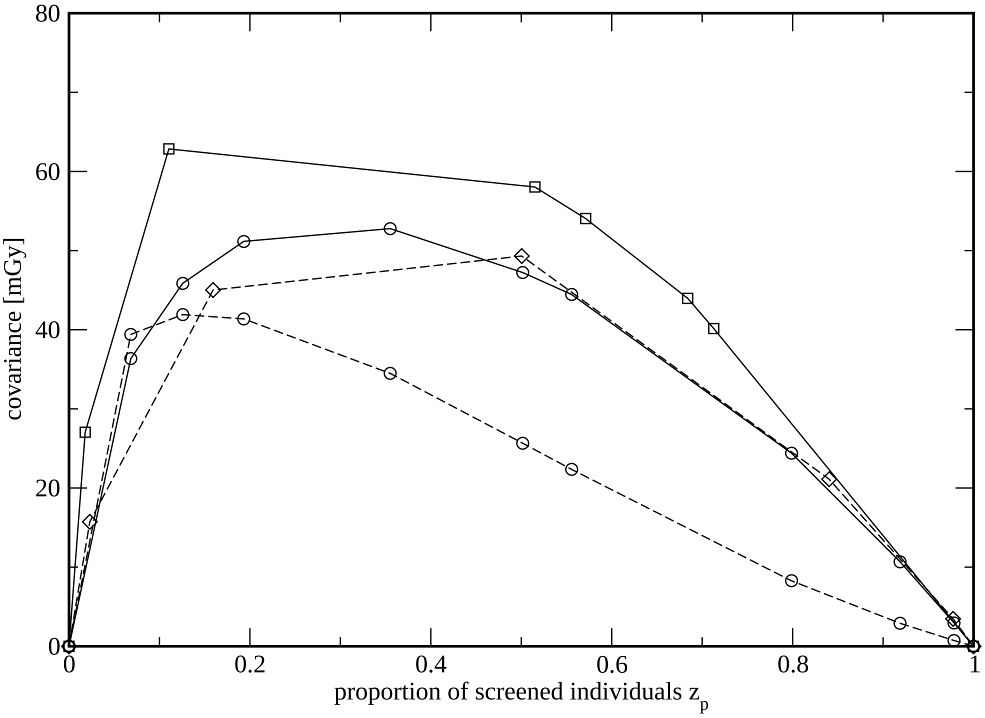


Figure 3. Simulated data points for the covariance between dose D and screening factor κ with κ_{ind} = 1, mean intra settlement covariance $\langle \text{cov}_I \rangle$ (– – –) and inter settlement covariance cov_S (——) for the percentile (◇), settlement (□), and individual (○) screening model, lines are only guides to the eye

4.3. Simulation of health status and Poisson regression

After each individual has been assigned a dose and a screening factor the tumour risk is fully defined and the health status can be determined. We compute the survival probability $\Psi_{ij} = \exp\left(-\Delta T\, h_{ij}\right)$ from the risk model of Table 1 and compare it with a random number Ψ_r which is evenly distributed between 0 and 1. If $\Psi_{ij} \leq \Psi_r$, a tumour case is assigned to the individual. Otherwise the tumour would appear after the end of the observation period. Competing risks have been neglected because they are very small for persons younger than 32 years of age. After these three simulation steps of Figure 1 a population data set for 366397 individuals is complete. The input parameter for the simulations are summarised in Table 1. For the situation without screening ($\kappa_{ind} = 0$) approx. 160 cases are generated which corresponds to the number of actually recorded cases in the study population. Thus the number of cases exceeds the recorded number when we introduce screening in our simulations.

To prepare the data for Poisson regression it must be stratified (or aggregated) on a settlement level. The means for dose and screening factor in each of the 743 settlements have to be calculated.

The number of expected cases in a settlement i is $\lambda_i = p_i \overline{h}_{i,eco}(h_{0,eco}, \beta_{eco})$ with the mean ecologic settlement risk from Equation (4) weighted by the person years $p_i = \Delta T N_i$.

Table 1. Simulation input parameter.

Name	Symbol	Value
EARPD	β	$2 \cdot 10^{-4}$/Gy/PY
background risk	h_0	$1 \cdot 10^{-5}$/PY
screening factor	κ_{ind}	1 and 2
individual risk	$h_{ij} = (1+\kappa_{ind}\,\theta_{ij})\,(h_0+\beta D_{ij})$	

The parameters $h_{0,eco}$ and β_{eco} are estimated by minimising the (log-)likelihood function

$$\ln L_P = -2\sum_i \left(n_i - \lambda_i + n_i \ln \frac{\lambda_i}{n_i} \right) \tag{8}$$

where n_i cases have been simulated (or observed) in a settlement. The MINUIT package of CERNLIB (James 1994) is used for minimisation.

By going through all the steps of the simulation scheme from Figure 1, only one set of risk parameter estimates is produced. To improve the accuracy of the risk estimates, 100 population data sets were simulated with the same input parameters. The point estimate and error bar for each regression parameter is then obtained by averaging over 100 runs (Figure 4).

5. Results

The simulations were done for the individual, settlement and percentile screening models using Condition (7) for the screening factor. The results with $\kappa_{ind} = 2$ are pooled in the Tables C1, C2 and C3 for three series of runs with different screening dose levels D_{scrn} or proportions z_p of screened individuals, respectively.

In Figure 4 100 point estimates for the ecologic EARPD β_{eco} from Poisson regression on the number of simulated cases are shown for the settlement model with dose level $D_{scrn} = 50$ mGy and individual screening factor $\kappa_{ind} = 1$. The average over 100 simulation runs meets the numerically exact value very well. For comparison the mean population-based EARPD β_{pop} of Equation (5) is also given. In this example, the ecologic EARPD overestimates the mean population-based EARPD by a factor of 1.17.

In the percentile model the simulation results yielded no bias (Table C3). This has also been proven with an analytical calculation in Appendix B.2. The reason is the special choice of the settlement correlations, which are all proportional to the mean settlement dose.

Figure 5 compares the ecologic EARPD β_{eco} with the true EARPD β_{pop} as a function of the fraction z_p of screened individuals for $\kappa_{ind} = 1,2$. Only the curves for the individual model and the percentile model are shown. If $z_p = 0$ the incidence is not increased by screening and the 'naked' EARPD β can be estimated unbiased. For $z_p = 1$ the EARPD β_{eco} equals $\beta_{pop} = (1+\kappa_{ind})\beta$ because the covariances vanish.

For both models the EARPD β_{pop} increases monotonously with z_p. However, the values for β_{eco} reach maxima between $0.5 < z_p < 0.6$. These maxima do not coincide with the maximal covariances which appear for $z_p < 0.4$. They are formed by two contradicting trends

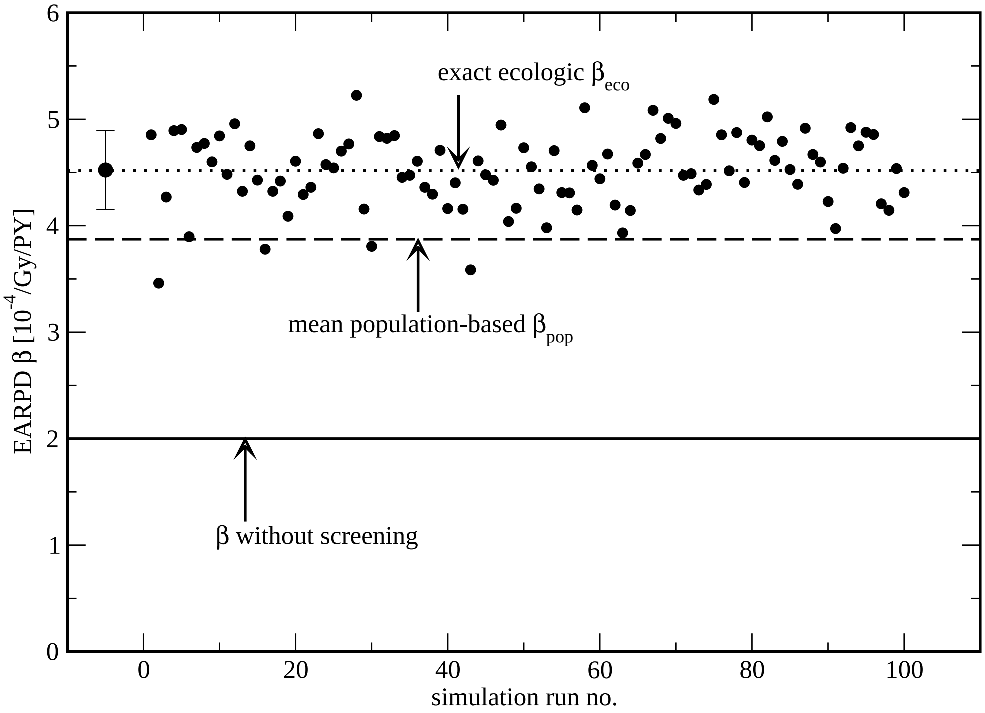


Figure 4. Point estimates (●) from Poisson regression on the number of simulated cases for the ecologic EARPD β_{eco} and the settlement screening model with dose level D_{scrn} = 50 mGy and individual screening factor κ_{ind} = 1, average over 100 points with standard error on the left, exact ecologic $\beta_{eco} = 4.52 \cdot 10^{-4}$/Gy/PY ($\cdots\cdots$) calculated from Equation (B.7); true mean population-based EARPD $\beta_{pop} = 3.87 \cdot 10^{-4}$/Gy/PY (– – –); $\beta = 2 \cdot 10^{-4}$/Gy/PY without screening (——) was simulation input

of falling covariances and rising mean screening factors $\langle\kappa\rangle$. The bias increases with rising κ_{ind}. It is higher in the settlement model than in the individual model.

The background risk $\langle h_0\rangle_{pop}$ is a linear function of the mean screening factor $\langle\kappa\rangle$ in the percentile model (Table C3). For $\kappa_{ind} = 1$ it remains always positive in the individual model and in the settlement model. But for $\kappa_{ind} = 2$ meaningless negative values can appear (Tables C1 and C2) which would lead to rejection of the underlying risk model in a real analysis.

With the Equations (B.2) and (B.7) we can calculate the relative bias numerically exact from the simulated population data. For all three screening models we found that the model-specific constant

$$C_M(z_p) = \frac{\beta_{eco} - \beta\,(1 + \langle\kappa\rangle)}{\langle\kappa\rangle\,\beta} \tag{9}$$

depends only on the proportion of screened individuals z_p but not on the screening factor κ_{ind}.

With this constant the relative bias for high values of κ_{ind} approaches the upper bound

$$\frac{\beta_{eco}}{\beta_{pop}} = \frac{1 + C_M(z_p)}{1 + \frac{\langle \mathrm{cov}_I(\kappa_{ind}=1)\rangle + \mathrm{cov}_S(\kappa_{ind}=1)}{z_p\langle D\rangle}}. \tag{10}$$

It acquires a maximum for small z_p. For z_p between 2-3% the bias is 1.6 for the individual model and 1.8 for the settlement model.

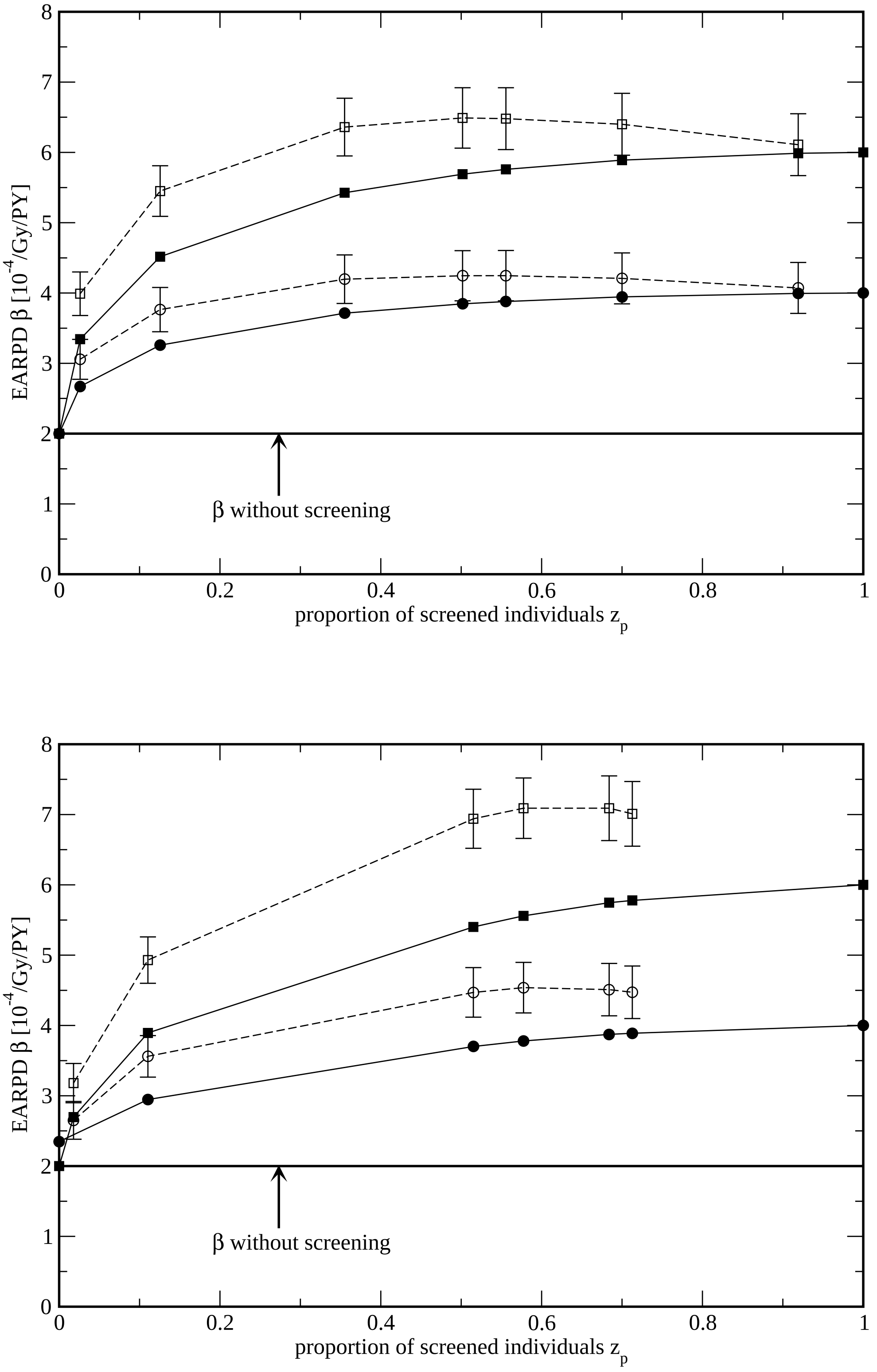


Figure 5. Comparison of the simulated ecologic EARPD β_{eco} (– – –) with standard errors and the true EARPD β_{pop} (——) for κ_{ind} = 1 (○ , ●) and κ_{ind} = 2 (□, ■) for the individual model (top) and the settlement model (bottom)

6. Discussion

We have investigated the effect of enhanced surveillance as a confounder on the risk estimate from ecologic studies. Based on aggregate post-Chernobyl data of radiation exposure to the thyroid of Ukrainian children, we have performed both Monte-Carlo simulations and analytical calculations. With a model of linear radiation-induced risk, which was combined with three different models for the correlation between exposure and screening, grouped population data with cancer cases have been generated. A fit on a settlement level with a linear risk model, that neglects screening information, produced an ecologic bias both for the background risk and the EARPD in most cases. However, with the percentile model we showed that the EARPD can be estimated unbiased even in the presence of correlations.

With analytical calculations the bias can be obtained in a simpler and more accurate way. We have shown that β_{eco} (and $h_{0,eco}$) can be calculated numerically by using the Equations (B.2) and (B.7). These equations establish the relations between the individual and ecologic risk parameters in the general case.

By comparing the ecologic risk parameter β_{eco} with the mean population-based risk parameter β_{pop} of Equation (5) we assessed the range of a possible bias. For our screening models we simulated a maximal factor of 1.3 for the relative bias of the EARPD. It occurs for intermediate values of z_p and for $\kappa_{ind} = 2$. In reality κ_{ind} can be much higher (Ron et al. 1992). In this case the relative bias increases asymptotically until it reaches an upper bound which stays well below the factor of two. This moderate value is partly caused by the assumption of a constant background risk. Uncertain dose estimates or migration from the original places of residence constitute additional potential for bias.

In reality the dose-screening correlations are expected to be weaker and will not follow strictly our didactic models. Moreover, screening scenarios are imaginable where the relative ecologic bias is smaller than one. This is the case if intensive screening took place in larger cities with low mean doses. Now screening and exposure are negatively correlated.

Finally, we can apply our methodology to assess the risk estimates for another hypothetical case. To date a cohort study with more than ten thousand members is carried out in the Ukraine (Tronko et al. 2003). Cases from this study will enter the cancer registry which will also be the data base of a future aggregate study. By mixing the cohort data and population-based aggregate data an additional bias will arise. We are able to estimate the order of magnitude of this bias by using Equations (B.2) and (B.7). We assume that 2.5% or 9160 children of our study population are included in this cohort. They are selected at random with a typical screening factor of 10. If approx. one half of the study population have been screened in conventional check-up programmes with a screening factor $\kappa_{ind} = 2$, the proportion z_p of screened individuals with the cohort included is only slightly above 50%. The EARPD $\beta_{m,eco}$ for the mixed cohort and population data can be compared with the EARPD β_{eco} without additional cohort screening. The relative difference $(\beta_{m,eco} - \beta_{eco})/\beta_{eco}$ lies between 5-11% for different screening scenarios as shown in Table 2. Because of these low percentage values this bias may be neglected.

Table 2. EARPD $\beta_{m,eco}$ from a study with mixed population-based and cohort data.

screening model	proportion z_p	EARPD $\beta_{m,eco}$ 10^{-4}/Gy/PY	rel. diff. in % $(\beta_{m,eco} - \beta_{eco})/\beta_{eco}$
random	0.5125	4.450	11
percentile	0.5125	5.555	8
individual	0.5141	6.889	6
settlement	0.5272	7.336	5

Acknowledgments

This work has been supported by the German Federal Minister of Environment, Nature Preservation and Reactor Safety and the German Federal Office of Radiation Protection under contract no. StSch 4299.

Appendix A. Mean risk for the total study population

Individual risk

For a person j from a settlement i, which has received a dose D_{ij}, we assume an individual risk

$$h_{ij} = (1+\eta_{ij})h_{0,ij} + (1+\kappa_{ij})\beta D_{ij}. \tag{A.1}$$

In general the individual screening factors η_{ij} and κ_{ij} could enhance the background risk $h_{0,ij}$ and the radiation-induced risk βD_{ij} with different magnitude.

Mean risk for one settlement

In a settlement i with N_i persons one obtains the mean risk (and all other settlement means)

$$\bar{h}_i = \frac{1}{N_i}\sum_j h_{ij} \tag{A.2}$$

by summing over all individuals j. The result is

$$\begin{aligned}\bar{h}_i &= \bar{h}_{0,i} + \frac{1}{N_i}\sum_j \eta_{ij}h_{0,ij} + \beta\bar{D}_i + \beta\frac{1}{N_i}\sum_j \kappa_{ij}D_{ij}\\ &= (1+\bar{\eta}_i)\,\bar{h}_{0,i}\left(1+\frac{\mathrm{cov}_{Ii}(\eta,h_0)}{(1+\bar{\eta}_i)\,h_{0,i}}\right) + (1+\bar{\kappa}_i)\,\beta\bar{D}_i\left(1+\frac{\mathrm{cov}_{Ii}(\kappa,D)}{(1+\bar{\kappa}_i)\,D_i}\right).\end{aligned} \tag{A.3}$$

The *intra* settlement covariance for exposure and screening (like the analog for background risk and screening) is defined as

$$\mathrm{cov}_{Ii}(\kappa,D) = \frac{1}{N_i}\sum_j \kappa_{ij}D_{ij} - \bar{\kappa}_i\bar{D}_i. \tag{A.4}$$

Mean risk for the total study population

To the mean risk for the total study population

$$\langle h\rangle = \frac{1}{p_{tot}}\sum_i p_i\bar{h}_i \tag{A.5}$$

each settlement i contributes its own mean risk $\bar{h}_i$. It is weighted by the person years $p_i = \Delta T N_i$ with the observation time ΔT, the total person years are $p_{tot} = \sum p_i$. By summing over all settlements i of Equation (A.3) one obtains the total mean risk

$$\begin{aligned}\langle h\rangle &= \langle h_0\rangle + \langle \mathrm{cov}_I(\eta,h_0)\rangle + \frac{1}{p_{tot}}\sum_i p_i\bar{\eta}_i\bar{h}_{0,i}\\ &+ \beta\,\langle D\rangle + \beta\,\langle \mathrm{cov}_I(\kappa,D)\rangle + \beta\,\frac{1}{p_{tot}}\sum_i p_i\bar{\kappa}_i\bar{D}_i\\ &= \langle h_0\rangle_{pop} + \beta_{pop}\,\langle D\rangle\,.\end{aligned} \tag{A.6}$$

with the true mean population-based background risk

$$\langle h_0\rangle_{pop} = (1+\langle\eta\rangle)\,\langle h_0\rangle\left(1+\frac{\langle \mathrm{cov}_I(\eta,h_0)\rangle + \mathrm{cov}_S(\bar{\eta},\bar{h}_0)}{(1+\langle\eta\rangle)\,\langle h_0\rangle}\right) \tag{A.7}$$

and the true mean EARPD

$$\beta_{pop} = (1 + \langle\kappa\rangle)\,\beta\left(1 + \frac{\langle \mathrm{cov}_I(\kappa, D)\rangle + \mathrm{cov}_S(\bar{\kappa}, \bar{D})}{(1 + \langle\kappa\rangle)\,\langle D\rangle}\right) \tag{A.8}$$

in the population. The *inter* settlement covariance

$$\mathrm{cov}_S(\bar{\kappa}, \bar{D}) = \frac{1}{p_{tot}} \sum_i p_i \bar{\kappa}_i \bar{D}_i - \langle\kappa\rangle\,\langle D\rangle \tag{A.9}$$

stems from the correlation between exposure and screening on a settlement level. An analogous definition holds for the inter settlement covariance between background risk and screening.

Appendix B. Two equations for the two risk parameters $h_{0,eco}$ and β_{eco}

Appendix B.1. General case

In ecologic analyses without screening information the settlement risk

$$\bar{h}_{i,eco} = h_{0,eco} + \beta_{eco} \bar{D}_i \tag{B.1}$$

is regressed on the settlement dose $\bar{D}_i$ to obtain the ecologic parameters $h_{0,eco}$ and β_{eco} for background risk and EARPD, respectively.

For the total study population the mean ecologic risk should equal the exact mean risk of Equation (A.6), i.e. $\langle h_{eco}\rangle = \langle h\rangle$. This condition yields the equation

$$h_{0,eco} + \beta_{eco}\,\langle D\rangle = \langle h_0\rangle_{pop} + \beta_{pop}\,\langle D\rangle\,. \tag{B.2}$$

for the risk parameters from ecologic regression and the true population-based parameters which use the full screening information.

To derive a second equation for $h_{0,eco}$ and β_{eco} we apply the (log-)likelihood function of Poisson regression

$$\ln L_P = -2 \sum_i \left(n_i - \lambda_i + n_i \ln\frac{\lambda_i}{n_i}\right) \tag{B.3}$$

with

$$n_i = p_i \bar{h}_i(\bar{h}_{0,i}, \beta) \quad \text{and} \quad \lambda_i = p_i \bar{h}_{i,eco}(h_{0,eco}, \beta_{eco}). \tag{B.4}$$

Normally n_i denotes the number of observed cases, but now we take this number directly from Equation (A.3). The expected cases λ_i are calculated from the ecologic risk $\bar{h}_{i,eco}$.

One can obtain a second equation by demanding that the derivative of L_P with respect to β_{eco}

$$-\frac{1}{2}\frac{\partial}{\partial\beta_{eco}} \ln L_P(\beta_{eco}) = \sum_i \left(\frac{n_i}{\lambda_i} - 1\right) \frac{\partial}{\partial\beta_{eco}}\lambda_i = 0. \tag{B.5}$$

For constant β and $\bar{h}_{0,i}$

$$\frac{\partial}{\partial\beta_{eco}}\lambda_i = p_i\,(\bar{D}_i - \langle D\rangle) \tag{B.6}$$

is a settlement-based derivative. The population-based mean of these derivatives vanishes. With the explicit expressions of $\overline{h}_i$ (A.3) and $\overline{h}_{i,eco}$ (B.1) one gets

$$\sum_i p_i \frac{(1+\overline{\eta}_i)\,\overline{h}_{0,i} + \mathrm{cov}_{Ii}\,(\eta, h_0) + \beta\,((1+\overline{\kappa}_i)\,\overline{D}_i + \mathrm{cov}_{Ii}\,(\kappa, D))}{h_{0,eco} + \beta_{eco} D_i} \\ \times\ (\overline{D}_i - \langle D \rangle) = 0. \tag{B.7}$$

We have now established two Equations (B.2) and (B.7) for the ecologic risk parameters β_{eco} and $h_{0,eco}$. Thus, an exact numerical calculation of the relative ecologic bias β_{eco}/β_{pop} (or $h_{0,eco}/\left\langle h_0 \right\rangle_{pop}$) from the population data is possible. One has to replace $h_{0,eco}$ with the help of Equation (B.2). Then Equation (B.7) can be solved numerically for β_{eco} with a routine for root finding like `rtbis ()` from the Numerical Recipes program library (Press et al. 1992).

Equation (B.7) contains implicitly various correlations like those between exposure and background risk. There exist even more complicated three-point-correlations between screening, exposure and background risk. They have not been taken into account in the definitions (A.7) and (A.8) of the true mean population-based risk factors. Therefore, we will obtain $\beta_{eco} \neq \beta_{pop}$ in the general case.

Appendix B.2. Special cases

For special cases without any bias we have found analytical expressions for β_{eco} which are discussed below.

Random screening. If all persons are selected at random for screening the mean settlement screening factors are all equal, i.e. $\overline{\eta}_i = \langle \eta \rangle$ and $\overline{\kappa}_i = \langle \kappa \rangle$ for all settlements i. Random selection suppresses all correlations between screening and exposure or screening and background risk, i.e. $\mathrm{cov}_{Ii} = 0$ and $\mathrm{cov}_S = 0$. By demanding $\overline{h}_{0,i} = h_0$ the correlation between exposure and background risk vanishes. Now Equation (B.7) is solved by

$$h_{0,eco} = (1+\langle \eta \rangle)h_0 \quad \text{and} \quad \beta_{eco} = (1+\langle \kappa \rangle)\beta. \tag{B.8}$$

Equal mean settlement screening factors with intra settlement correlations. This situation is treated by the percentile model for constant background risk h_0, where in each settlement the same percentile z_p of children with higher doses are screened. Hence, intra settlement correlations $\mathrm{cov}_{Ii} \neq 0$ will arise. But they are proportional to $\kappa_{ind}\overline{D}_i$. The constant of proportionality depends on z_p but is the same for each settlement. As a consequence, all mean settlement screening factors are equal and the inter settlement covariance $\mathrm{cov}_S(\overline{\kappa}, \overline{D})$ vanishes. Now the relations

$$h_{0,eco} = (1+\langle \eta \rangle)h_0 \quad \text{and} \quad \beta_{eco} = (1+\langle \kappa \rangle)\beta \left(1 + \frac{\langle \mathrm{cov}_I \rangle}{(1+\langle \kappa \rangle)\,\langle D \rangle}\right) \tag{B.9}$$

solve Equation (B.7).

Appendix C. Tables with simulation results

All values for the covariances, the cases, and the risk parameters with errors are the averages over 100 simulation runs. The error bars for the risk parameters are the standard errors calculated from the parabolic approximation of the likelihood function, i.e. Wald-based standard errors. For the simulated number of cases n_c the error bars are calculated from the standard deviation over 100 runs. They are consistent with the theoretically expected value of $\sqrt{n_c}$.

Table C1. Individual screening model with $\kappa_{ind} = 2$, values for $h_{0,eco}$ and β_{eco} from Poisson regression on the number of simulated cases and exact numerical solution of Equations (B.2) and (B.7).

pro-	scrn.	dose lev.	covariances		simul.	backgrd. risk $h_{0,eco}$		EARPD β_{eco}		β_{pop}	rel. eco.
portion	factor	D_{scrn}	$\langle \mathrm{cov}_I \rangle$	cov_S	cases	10^{-6}/PY		10^{-4}/Gy/PY			bias
z_p	$1 + \langle \kappa \rangle$	mGy	mGy		n_c	regression	exact	regression	exact	Eq. (A.8)	β_{eco}/β_{pop}
0.0261	1.052	1000	63.0	44.2	248 ± 14	-1.13 ± 3.06	-1.435	3.99 ± 0.31	4.03	3.34	1.21
0.1254	1.251	300	83.3	91.1	330 ± 16	-4.22 ± 3.51	-4.377	5.45 ± 0.36	5.49	4.52	1.22
0.3550	1.710	100	68.7	104.9	404 ± 18	0.15 ± 4.59	0.213	6.36 ± 0.41	6.40	5.43	1.18
0.5017	2.003	60	51.3	94.5	432 ± 18	5.63 ± 5.16	5.779	6.49 ± 0.43	6.51	5.69	1.14
0.5556	2.111	50	44.6	88.4	441 ± 19	8.00 ± 5.33	8.047	6.48 ± 0.44	6.51	5.76	1.13
0.7000	2.399	30	27.2	67.4	459 ± 18	14.5 ± 5.7	14.69	6.40 ± 0.44	6.43	5.89	1.09
0.9191	2.838	10	5.8	21.2	481 ± 19	25.4 ± 6.1	25.72	6.11 ± 0.44	6.14	5.99	1.03

Table C2. Settlement screening model with κ_{ind} = 2, values for $h_{0,eco}$ and β_{eco} from Poisson regression on the number of simulated cases and exact numerical solution of Equations (B.2) and (B.7).

pro-	scrn.	dose lev.	covar.	simul.	backgrd. risk $h_{0,eco}$		EARPD β_{eco}		β_{pop}	rel. eco.
portion	factor	D_{scrn}	cov_S	cases	10^{-6}/PY		10^{-4}/Gy/PY			bias
z_p	$1+\langle\kappa\rangle$	mGy	mGy	n_c	regression	exact	regression	exact	Eq. (A.8)	β_{eco}/β_{pop}
0.0179	1.036	1000	54.1	208 ± 14	1.73 ± 3.00	1.736	3.18 ± 0.28	3.19	2.70	1.18
0.1104	1.221	300	125.7	291 ± 15	-5.77 ± 2.82	-5.937	4.93 ± 0.33	4.94	3.89	1.27
0.5151	2.030	130	116.1	416 ± 18	-6.79 ± 4.24	-6.806	6.94 ± 0.42	6.97	5.40	1.29
0.5775	2.155	100	108.1	429 ± 18	-5.55 ± 4.59	-5.448	7.09 ± 0.43	7.12	5.56	1.28
0.6840	2.368	50	87.5	449 ± 18	-0.24 ± 5.38	-0.021	7.09 ± 0.46	7.12	5.75	1.24
0.7128	2.426	40	80.3	453 ± 18	2.22 ± 5.56	2.553	7.01 ± 0.46	7.04	5.78	1.22

Table C3. Percentile screening model with κ_{ind} = 2, values for $h_{0,eco}$ and β_{eco} from Poisson regression on the number of simulated cases and from the Formulas (B.9).

per-	scrn.	covar.	simul.	backgrd. risk $h_{0,eco}$		EARPD β_{eco}		rel. eco.
centile	factor	$\langle \mathrm{cov}_I \rangle$	cases	10^{-6}/PY		10^{-4}/Gy/PY		bias
z_p	$1+\langle\kappa\rangle$	mGy	n_c	regression	Eq. (B.9)	regression	Eq. (B.9)	β_{eco}/β_{pop}
0.0228	1.046	31.0	193 ± 13	10.4 ± 3.8	10.46	2.44 ± 0.26	2.45	1
0.1587	1.317	88.6	279 ± 15	12.9 ± 4.5	13.17	3.65 ± 0.34	3.66	1
0.5	2	98.1	396 ± 19	19.7 ± 5.4	20.00	5.11 ± 0.40	5.13	1
0.8413	2.683	42.0	466 ± 19	26.3 ± 6.0	26.83	5.83 ± 0.43	5.85	1
0.9772	2.954	6.9	485 ± 19	29.0 ± 6.2	29.54	5.97 ± 0.44	5.99	1

References

Astakova, L. N., Anspaugh, L. R., Beebe, G. W., Bouville, A., Drozdovitch, V. V., Garber, V., Gavrilin, Y. I., Khrouch, V. T., Kuvshinnikov, A. V., Kuzmenkov, Y. N., Minenko, V. P., Moschik, K. F., Nalivko, A. S., Robbins, J., Shemiakina, E. V., Shinkarev, S., Tochitskaya, S. I., and Waclawiw, M. A. (1998). Chernobyl-related thyroid cancer in children of Belarus: a case-control study. *Radiation Research* **150**, 349–356.

Buglova, E. E., Kenigsberg, J. E., and Golovneva, A. (1996). Cancer risk estimation in Belarussian children due to thyroid irradiation as a consequence of the Chernobyl nuclear accident. *Health Physics* **71**, 45–49.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review* **43**, 557–572.

Greenland, S. (2001). Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology* **30**, 1343–1350.

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine* **21**, 389–395.

Jacob, P., Goulko, G., Heidenreich, W. F., Likhtarev, I., Kairo, I., Tronko, N. D., Bogdanova, T. I., Kenigsberg, J., Buglova, E., Drozdovitch, V., Golovneva, A., Demidchik, E. P., Balanov, M., Zvonova, I., and Beral, V. (1998). Thyroid cancer risk to children calculated. *Nature* **392**, 31–32.

Jacob, P., Kenigsberg, Y., Zvonova, I., Goulko, G., Buglova, E., Heidenreich, W. F., Golovneva, A., Bratilova, A. A., Drozdovitch, V., Kruk, J., Pochtennaja, G. T., Balanov, M., Demidchik, E. P., and Paretzke, H. G. (1999). Childhood exposure due to the Chernobyl accident and thyroid cancer risk in contaminated areas of Belarus and Russia. *British Journal of Cancer* **80**(9), 1461–1469.

James, F. (1994). *MINUIT - Function minimization and error analysis, version 94.1*. CERN, Geneva. CERN Program Library Entry D506.

Likhtarev, I. A., Kayro, I. A., Shpak, V. M., Tronko, N. D., and Bogdanova, T. I. (1999). Radiation-induced and background thyroid cancer of Ukrainian children (Dosimetric approach). *International Journal of Radiation Medicine* **3–4**, 51–66.

Lubin, J. H. (1998). On the discrepancy between epidemiologic studies in individuals of lung cancer and residential radon and Cohen’s ecologic regression. *Health Physics* **75**(1), 4–20.

Lubin, J. H. (2002). The potential for bias in Cohen’s ecological analysis of lung cancer and residential radon. *Journal of Radiological Protection* **22**, 141–148.

Morgenstern, H. (1998). Ecologic studies. *Modern Epidemiology* (eds. Rothman, K. J. and Greenland, S.), chap. 23, 459–480. Lippincott-Raven, Philadelphia.

NCRP (2001). Evaluation of the Linear-Nonthreshold Dose-Response Model for Ionizing Radiation. Tech. Rep. 136, National Council on Radiation and Measurements, Bethesda, Maryland.

Piantadosi, S. (1994). Invited commentary: Ecologic Biases. *American Journal of Epidemiology* **139**(8), 761–764.

Piantadosi, S., Byar, D. P., and Green, S. B. (1988). The ecologic fallacy. *American Journal of Epidemiology* **127**, 893–904.

Press, W. H., Flannery, B. P., Teukolski, S. A., and Vetterling, W. T. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge, MA, 2nd edn.

Ron, E., Lubin, J., and Schneider, A. B. (1992). Thyroid cancer incidence. *Nature* **360**, 113.

Stidley, C. A. and Samet, J. M. (1994). Assessment of Ecologic Regression in the Study of Lung Cancer and Indoor Radon. *American Journal of Epidemiology* **139**(3), 312–322.

Tronko, M. D., Bobylyova, O. O., Bogdanova, T. I., Epshtein, O. V., Likhtaryov, I. A., Markov, V. V., Oliynyk, V. A., Tereshchenko, V. P., Shpak, V. M., Beebe, G., Bouville, A., Brill, A., Burch, D., Fink, D., Greenebaum, E., Howe, G., Luckyanov, N., Masnyk, I., McConnell, R., Robbins, J., Thomas, T., and Voillequé, P. (2003). Thyroid gland and radiation (Ukrainian-American thyroid project). *Radiation and Humankind* (eds. Shibata, Y., Yamashita, S., Watanabe, M., and Tomonaga, M.), vol. 1258 of *International Congress Series*, 91–104. Elsevier, Amsterdam. Proceedings of the 1st Nagasaki Symposium of the International Consortium for Medical Care of Hibakusha and Radiation Life Science, Nagasaki, Japan, 21-22 February 2003.

USSR (1991). The USSR 1989 census: The age-by-sex population distribution of the UkSSR. The UkSSR Ministry of Statistics Publication. Kiev.