Supplementary Material

# Integration of a radiation biomarker into modelling of thyroid carcinogenesis and post-Chernobyl risk assessment

J.C. Kaiser, R. Meckbach, M. Eidemüller,
M. Selmansberger, K. Unger, V. Shpak,
M. Blettner, H. Zitzelsberger and P. Jacob

August 10, 2016

# Materials and Methods

## Data sets

### UkrAm cohort for radiation epidemiology

After a transfer agreement had been signed between Institute of Endocrinology and Metabolism of the National Academy of Medical Sciences, Ukraine and the National Cancer Institute, Bethesda (MD), USA as data providers and the Helmholtz Zentrum München, Germany as recipient investigator a dedicated data set of the UkrAm cohort was made available in April 2013. It contained 13,243 subjects who took part in the first screening cycle 1998 - 2000. Tronko et al. [1] excluded 116 subjects for various reasons and used 13,127 subjects for their analysis. Brenner et al. [2] revised this number further downwards to 12,514 subjects mainly due to loss of follow-up.

In the present study the date of accident on 26. April 1986 was chosen as begin of follow-up for all subjects. End of follow-up for cohort members diseased with PTC was set to date of operation. For healthy subjects is was fixed to the end of 2008 to include the last date of operation. Calendar dates for subjects contained month and year. The exact day was not available and has been replaced by the 15th day of each month. In principle, the prevalent cases from the first screening cycle listed in Table S1 could have occurred before the accident. However, since thyroid cancer in early childhood is a very rare disease early occurrence is very unlikely. With the modeled baseline rate for the present data set one can estimate that no case occurred before the accident.

In the present study 57 subjects were discarded due to missing dose estimates, exposure in utero or at age $\geq$ 18 yr, and date of birth after April 1986. Another 3 subjects with diagnosed thyroid cancer but missing histological classification have also been discarded. The level of detail in the present data set of UkrAm cohort did not allow to completely reproduce the designs of the previous studies [1, 2]. The data set of the remaining 13,183 subjects is summarized in Table S2.

Table S1: Thyroid cancer cases by screening period and histological type.

| screening period | case type in Eq. (S17) | papillary | follicular | medullary | all types |
|---|---|---|---|---|---|
| before 1998 | prevalent | 11 | 0 | 0 | 11 |
| $1^{st}$: 1998 - 2000 | prevalent | 43 | 2 | 0 | [a]45 |
| $2^{nd}$ - $4^{th}$: 2001 - 2008 | incidental | [b]61 | 3 | 1 | [b]65 |
| sum | | [c]115 | 5 | 1 | 121 |

[a]cases used in Tronko et al. [1]
[b]cases used in Brenner et al. [2]
[c]cases used in present study

3

Table S2: Summary of the epidemiological UkrAm cohort with arithmetic means for age at exposure (AaE), time since exposure (TsE), age at operation (AaO) and thyroid dose.

| | unit | both sexes | male | female |
|---|---|---|---|---|
| subjects | - | 13,183 | 6,474 | 6,678 |
| person years | PY | 298,463 | 146,992 | 151,471 |
| AaE (min; max) | yr | 7.8 (0.0; 17.9) | 7.8 (0.0; 17.9) | 7.8 (0.0; 17.9) |
| TsE of PTC cases (min; max) | yr | 16.3 (3.9; 22.5) | 16.3 (3.9; 22.5) | 16.2 (4.6; 22.1) |
| AaO of PTC cases (min; max) | yr | 24.4 (7.4; 35.1) | 24.9 (7.4; 35.1) | 24.1 (8.5; 34.1) |
| thyroid dose (geom. mean) | Gy | 0.663 (0.219) | 0.690 (0.235) | 0.637 (0.205) |
| thyroid dose of PTC cases (geom. mean) | Gy | 1.42 (0.447) | 1.13 (0.331) | 1.59 (0.578) |
| number of PTC cases | - | 115 | 44 | 71 |

# Mechanistic model of PTC pathogenesis

The mechanistic model of the present study is an instantiation from a class of cancer induction models in the mathematical framework of Little and colleagues [3,4]. They have generalized the two-stage clonal expansion (TSCE) model of carcinogenesis proposed by Moolgavkar and Knudson [5]. The TSCE model relies on two rate-limiting cell transitions which are separated by clonal expansion of initiated cells. Transition rates pertain to the (yearly) occurrence of key events on the pathway to cancer. In a process-oriented modelling approach these events are related to the development of various forms of genetic damage such as mutations, hypermethylation, gene rearrangements or instabilities in whole chromosomes. Although spatial expansion processes play a role in carcinogenesis, the present class of mechanistic models is only concerned with time dependencies. Transition rates and rates of cell division or inactivation are treated as transient Poisson point processes of cell birth and death which are expressed in a set of coupled master equations [6]. These equations describe the transient evolution of numbers of cells which have accumulated genetic instability in a sequence of stages. In principle, numbers and properties of such cells should be accessible by experimental investigation of each stage.

In the present study the so-called deterministic model version is applied which considers only *expectation values* of cell numbers. In the stochastic version statistical fluctuations in cell numbers are fully included [7,8]. Such fluctuations influence the cancer incidence rate mainly at older age. For the young UkrAm cohort they can be neglected. This has been tested by fitting both stochastic and deterministic models. Model parameters related to fluctuations in cell numbers did not improve the fit.

### Two stage model with radiation action on first stage

The two stage model of Figure S1 describes radiation-induced CLIP2-associated carcinogenesis (rC2C). It starts with a very large number of $N_b$ susceptible cells (i.e. thyroblasts) in homeostasis which undergo two transitions $\mu_0(a)$ and $\mu_1$ to a (radiation-induced) PTC. The second stage of a detectable PTC is reached from the first stage by asymmetric cell division. The master equations for the mean values of cell numbers $N_1(a)$ and $N_2(a)$ at attained age $a$ in the first and second stage are

$$\frac{\mathrm{d}}{\mathrm{d}a}N_1(a) \;=\; N_b\,\mu_0(a) \quad \text{with initial condition} \quad N_1(0) = 0 \qquad \text{(S1)}$$

$$\frac{\mathrm{d}}{\mathrm{d}a}N_2(a) \;=\; \mu_1\,N_1(a) \quad \text{with initial condition} \quad N_2(0) = 0.$$

For constant coefficients the master equations are easily integrated with boundaries at $0$ and $a$ to yield

$$h_{rC2C}(a) \;=\; \mu_1\,N_1(a) = X_0\,a \qquad \text{(S2)}$$

$$\ln S_{rC2C}(a) \;=\; -N_2(a) = -\frac{1}{2}X_0\,a^2.$$

for hazard $h_{rC2C}(a)$ and survival function $S_{rC2C}(a)$. $N_2(a)$ represents the expectation value of a Poisson-distributed random variable pertaining to the number of cells in the last box of the two stage model in Figure S1. For a Poisson distribution in general, an expectation value is given by $Np$ where $p$ denotes the small probability of developing a malignant cell from a large number of $N$ precursor cells. In this case we get $S = (1-p)^N \simeq \exp(-Np)$ and $N_2(a) = Np$.

$X_0 = N_b\mu_0\mu_1$ is called the Armitage-Doll parameter. The term 'Armitage-Doll parameter' appeals to the product of transition rates which determine the incidence rate in the model of Armitage and Doll [9].

Exposure increases the first transition rate

$$\mu_0(a) = \begin{cases} \mu_0 & : & a < a_e \\ \mu_0 + \mu_r & : & a_e \le a < a_e + \Delta t_e \\ \mu_0 & : & a \ge a_e + \Delta t_e \end{cases} \qquad \text{(S3)}$$

by $\mu_r$ for a short period $\Delta t_e = 7/365$ yr of one week after age at exposure $a_e$. Before and after exposure the first transition rate $\mu_0$ remains constant.

For short exposure one obtains

$$N_1(a) \;=\; N_b\mu_0 a + \Theta(a - a_e)N_b\mu_r\Delta t_e \qquad \text{(S4)}$$

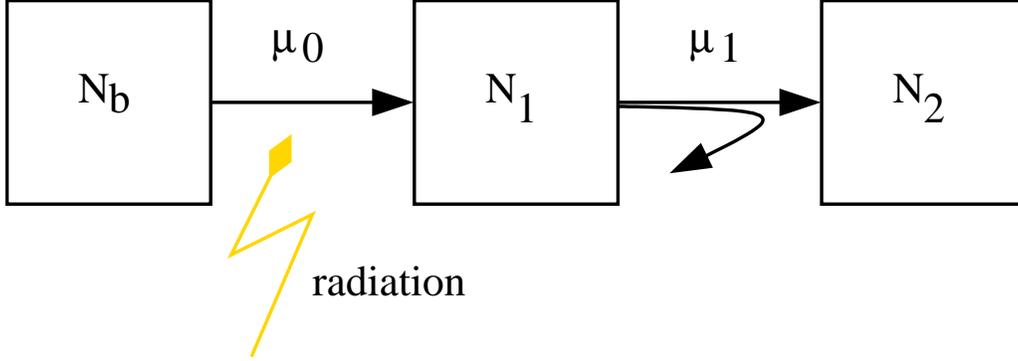$$N_2(a) \;=\; \frac{1}{2}X_0 a^2 + \Theta(a - a_e)X_0\frac{\mu_r}{\mu_0}\Delta t_e\,a$$

and with Eq. (S2)

Figure S1: Two stage model with radiation action; radiation enhances the first transition rate $\mu_0 \to \mu_0 + \mu_r$ for a short period $\Delta t_e = 7/365$ yr of one week.

$$h_{rC2C}(a) = X_0\, a + \Theta(a - a_e)\, X_0\, \frac{\mu_r}{\mu_0} \Delta t_e \tag{S5}$$

$$\ln S_{rC2C}(a) = -\frac{1}{2} X_0\, a^2 - \Theta(a - a_e)\, X_0\, \frac{\mu_r}{\mu_0} \Delta t_e\, (a - a_e).$$

The Heaviside function $\Theta(a - a_e)$ equals one if attained age $a$ exceeds age at exposure $a_e$ and zero otherwise. The radiation response

$$X_0\, \frac{\mu_r}{\mu_0} = \kappa\, D \exp(\epsilon D + \sigma s) \tag{S6}$$

increases linearly with coefficient $\kappa$ for dose $D$ and is attenuated exponentially with coefficient $\epsilon$ for higher doses. The strong dependence on sex $s = -1, +1$ for males, females is modelled with coefficient $\sigma$.

## Two stage model with age-dependent second stage

To explain the occurrence of sporadic PTC with a positive CLIP2 marker (sporadic CLIP2-associated carcinogenesis or sC2C) in patients operated above age 20 an additional path has been added to the mechanistic model. For this path the second stage of the two stage model has been modified. The second transition rate is increased with the step function

$$\mu_1(a) = \frac{\hat{\mu}_1}{2} \left[ \tanh\left( s_{\mu_1}(a - a_{\mu_1}) \right) + 1 \right]. \tag{S7}$$

The step has a slope $s_{\mu_1}$ and is centered around age $a_{\mu_1}$. For old ages the second transition rate approaches $\hat{\mu}_1$.

The hazard function $h_{sC2C}(a)$ has a closed form and the corresponding survival function $S_{sC2C}(a)$ can be derived by numerical integration according to

$$h_{sC2C}(a) = \frac{\hat{X}_0}{2} \left[ \tanh \left( s_{\mu_1}(a - a_{\mu_1}) \right) + 1 \right] a \tag{S8}$$

$$\ln S_{sC2C}(a) = -\frac{\hat{X}_0}{2} \int_0^a \left[ \tanh \left( s_{\mu_1}(a' - a_{\mu_1}) \right) + 1 \right] a' \mathrm{d}a'$$

with the Armitage-Doll parameter $\hat{X}_0 = N_b \mu_0 \hat{\mu}_1$.

## Three stage model with clonal expansion

The three stage model of Figure S2 starts with a larger number of $N_s$ thyrocytes which acquire two hits with transition rates $\nu_0$ and $\nu_1$ until they reach an intermediate stage. In this stage they can expand into clones with net growth rate $\gamma = \alpha - \beta$ if symmetric cell division $\alpha$ exceeds cell inactivation $\beta$. Transition to the cancerous third stage is accomplished with rate $\nu_2$. The corresponding master equations for the mean values of cell numbers $N_1(a)$, $N_2(a)$ and $N_3(a)$ are given by

$$\frac{\mathrm{d}}{\mathrm{d}a} N_1(a) = N_s \nu_0 \quad \text{with initial condition} \quad N_1(0) = 0 \tag{S9}$$

$$\frac{\mathrm{d}}{\mathrm{d}a} N_2(a) = \nu_1 N_1(a) + \gamma N_2(a) \quad \text{with initial condition} \quad N_2(0) = 0$$

$$\frac{\mathrm{d}}{\mathrm{d}a} N_3(a) = \nu_2 N_2(a) \quad \text{with initial condition} \quad N_3(0) = 0.$$

For constant coefficients integration of the master equations yields

$$N_1(a) = N_s \nu_0 a \tag{S10}$$

$$N_2(a) = \frac{N_s \nu_0 \nu_1}{\gamma^2} \left( \exp(\gamma a) - 1 - \gamma a \right)$$

$$N_3(a) = \frac{N_s \nu_0 \nu_1 \nu_2}{\gamma^3} \left( \exp(\gamma a) - 1 - \gamma a - \frac{1}{2}(\gamma a)^2 \right)$$
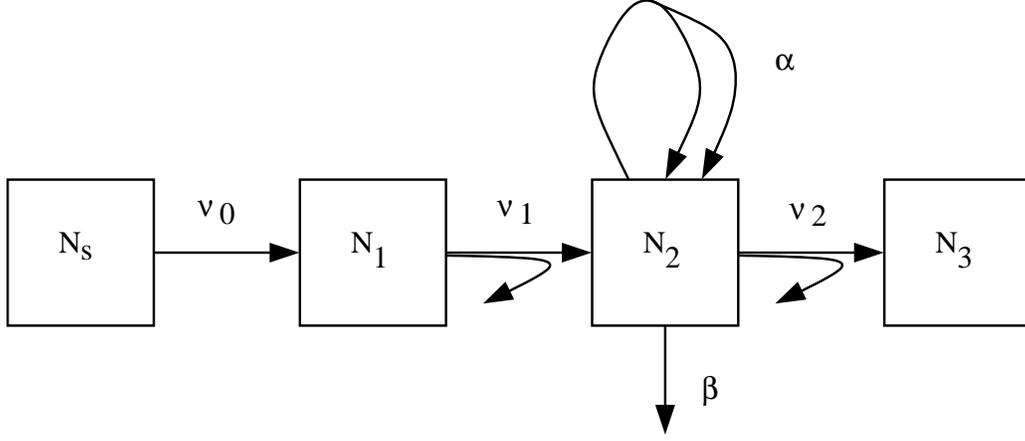
Figure S2: Three stage model with clonal expansion; in the second stage clones grow with a net promotion rate $\gamma = \alpha - \beta$ if symmetric cell division $\alpha$ exceeds cell inactivation $\beta$.

Using the Armitage-Doll parameter $R_0 = N_s \nu_0 \nu_1 \nu_2$ one obtains

$$
\begin{aligned}
h_{MSC}(a) &= \nu_2\, N_2(a) = \frac{R_0}{\gamma^2} \left( \exp(\gamma a) - 1 - \gamma a \right) & \text{(S11)} \\
\ln S_{MSC}(a) &= -N_3(a) = -\frac{R_0}{\gamma^3} \left( \exp(\gamma a) - 1 - \gamma a - \frac{1}{2}(\gamma a)^2 \right)
\end{aligned}
$$

for the sporadic hazard $h_{MSC}(a)$ and survival function $S_{MSC}(a)$ in the MSC path.

## Decomposition of hazard and survival function for the preferred mechanistic model

The total hazard function of the preferred mechanistic model is obtained by combining the two stage model with radiation action on the first stage (rC2C), the two stage model with increasing age dependence of the second stage (sC2C) and the three stage model with clonal expansion (MSC). The corresponding functions

$$h(a) = \underbrace{h_{rC2C}(a) + h_{sC2C}(a)}_{=h_{C2C}(a)} + h_{MSC}(a) \qquad \text{(S12)}$$

$$S(a) = \underbrace{S_{rC2C}(a) \, S_{sC2C}(a)}_{=S_{C2C}(a)} \, S_{MSC}(a).$$

are given by the expressions of Eqs. (S5),(S8) and (S11).

## Identifiability of model parameters

It is well known that not all biological parameters for this class of mechanistic models can be identified from the incidence data, although at least in principle they should be measurable in experiments [10]. The problem with parameter identifiability is caused by the mathematical model structure and cannot be removed by increasing statistical power. Only combinations of parameters can be identified which depend on the chosen model formulation. In the present analysis the Armitage-Doll parameter $R_0$ and the clonal expansion rate $\gamma$ can be estimated for the sporadic path . For the radiation-induced path the three dose response parameters $\kappa$, $\varepsilon$ and $\varsigma$ are identifiable. The Armitage-Doll parameter $X_0$ can be estimated in principle but comes out very small. It has been fixed to a value of $3.929 \times 10^{-11}\,\text{yr}^{-2}$ to prevent numerical instabilities and is not counted as adjustable parameter. The three parameters $\hat{X}_0$, $s_{\mu_1}$ and $a_{\mu_1}$ in Eq. (S8) for the sporadic submodel of sC2C have been fixed in the models fits.

## Descriptive risk model

A descriptive model for the excess absolute risk (EAR)

$$
\begin{aligned}
h(a) &= h_0(a) + \Theta(a - a_e)\,\text{EAR}(s, D) \quad \text{with} \quad & \text{(S13)} \\
h_0(a) &= 10^{-4}\text{PY}^{-1}\,\exp(b_0 + b_a \ln(a/25)) \\
\text{EAR}(s, D) &= \text{ear}\,D\exp(b_{exp}D + b_s s)
\end{aligned}
$$

has been applied to the epidemiological data to put the results from the mechanistic model into perspective. The dose response depends on sex $s$ similar to the response (S6) of the mechanistic model. ERR models have also been applied but yielded fits of inferior quality if dose effect modifiers were applied in the same way as for EAR models.

## Oblast adjustment

Although the mean dose for members exposed in Zhytomir oblast is higher by factors 3-4 compared to oblasts Kyiv and Chernihiv, oblast-specific incidence rates are similar in view of their uncertainties (Table S3). This mismatch has been attributed to unmeasured explanatory variables such as iodine deficiency [2]. Both descriptive and mechanistic models have been adjusted for these unexplained variations by a heuristic factor $\exp(\Phi_{obl86})$ pertaining to the total risk according to $h \rightarrow \exp(\Phi_{obl86})\,h$ for residents of Zhytomir oblast in 1986. Estimates of $\Phi_{obl86}$ for the preferred models are given in Tables S5 and S6.

Table S3: Persons, PTC cases, incidence rates per 10,000 person years (PY) with 95% CI in brackets, mean age at exposure (AaE), mean age at operation of case (AaO) and mean thyroid dose in Zhytomir, Kyiv and Chernihiv oblasts of residence in 1986.

| oblast | persons | cases | incidence per $10^4$ PY (95% CI) | mean AaE (yr) | mean AaO of cases (yr) | mean dose (Gy) |
|---|---|---|---|---|---|---|
| Zhytomir | 3,659 | 28 | 4.0 (3.1; 5.0) | 7.9 | 24.9 | 1.411 |
| Kyiv | 2,595 | 24 | 4.1 (2.6; 5.8) | 9.0 | 25.1 | 0.493 |
| Chernihiv | 6,929 | 63 | 3.4 (2.2; 4.7) | 7.3 | 24.0 | 0.322 |
| total | 13,183 | 115 | 3.9 (3.2; 4.6) | 7.8 | 24.4 | 0.663 |

## Statistical analysis

### Individual likelihood regression

The method of individual likelihood regression directly accounts for the health status of all cohort members. In the likelihood function the individual status are expressed by probabilities of dedicated mathematical forms which are derived below. The method accommodates for differential information on the circumstances of tumor detection in subsequent screening cycles. 13,037 cohort members were free of PTC between begin of follow-up at age $a_b$ and end of follow-up at age $a_e$ (Table S1). 54 PTCs were operated at age $a_o$ after detection before or during the 1st screening cycle. These cases are termed 'prevalent' based on the assumption that a tumor appeared at an *unknown* point in time between $a_b$ and $a_o$. 61 PTCs were detected during the 2nd to 4th screening cycle and are termed 'incidental'. They developed between the last examination with negative finding at age $a_s$ and age at operation $a_o$. The difference between prevalent and incidental cases is given by the observation that generally $\Delta a_{prev} = a_o - a_b \gg \Delta a_{inc} = a_o - a_s$.

The hazard function

$$
\begin{aligned}
h(a) &= -\frac{\mathrm{d}S(a)/\mathrm{d}a}{S(a)} \\
&\simeq \frac{N_{cases}}{N_{coh}\Delta a}
\end{aligned}
\tag{S14}
$$

at attained age $a$ is defined by the negative derivative of the survival function $S$ taken at $a$ and divided by $S(a)$. The hazard is approximated by the incidence rate given by the number $N_{cases}$ of tumors among $N_{coh}$ cohort members observed in an age interval $\Delta a$ at risk. The survival function $S(a)$ denotes the probability for a cohort member to be tumor-free at age $a$. Regression analysis was performed with the log-likelihood function

$$
\ln \mathcal{L} = \sum_{\substack{\text{tumor-free} \\ \text{members}}} \ln p_{free} + \sum_{\substack{\text{prevalent} \\ \text{cases}}} \ln p_{prev} + \sum_{\substack{\text{incidental} \\ \text{cases}}} \ln p_{inc}
\tag{S15}
$$

where the individual probabilities $p_{free}$, $p_{prev}$ and $p_{inc}$ for cohort members are derived from

12

$$
\begin{aligned}
S_b\, p_{free} &= S_e &&: \text{no tumor in follow-up period} & \text{(S16)}\\
S_b\, p_{prev} &= S_b - S_o &&: \text{tumor detected before or during } 1^{st} \text{ screening}\\
S_b\, p_{inc} &= S_s - S_o \simeq h_o S_o \Delta a_{inc} &&: \text{tumor detected in } 2^{nd}\text{-}4^{th} \text{ screening}
\end{aligned}
$$

with notation $S_b = S(a_b)$ , $S_e = S(a_e)$, $S_s = S(a_s)$, $S_o = S(a_o)$, $h_o = h(a_o)$ and $\Delta a_{inc} \ll a_o$. The individual probabilities are conditioned on the probability $S_b$ of being tumor-free at begin of follow-up.

Inserting the dedicated probabilities of Eq. (S16) into $\ln \mathcal{L}$ of Eq. (S15) yields

$$
\ln \mathcal{L} = \sum_{\substack{\text{tumor-free}\\\text{members}}} \ln\left(S_e/S_b\right) + \sum_{\substack{\text{prevalent}\\\text{cases}}} \ln(1 - S_o/S_b) + \sum_{\substack{\text{incidental}\\\text{cases}}} \ln\left(h_o S_o/S_b\right) \quad \text{(S17)}
$$

where $\Delta a_{inc}$ for incidental cases has been omitted since it does not depend on model parameters [11].

Maximum likelihood estimates for model parameters are taken at the minimum of the deviance $D = -2 \min \ln \mathcal{L}$. The contributions to the deviance from tumor-free cohort members are negative but close to zero. Numerical values of prevalent cases range between -3.5 and -6.5, for incidental cases the values lie between -7.0 and -8.2. Hence, maximum likelihood estimates are predominantly determined by incidental cases and to a lesser extent by prevalent cases whereas the influence of tumor-free subjects is rather small.

## Parameter estimation and uncertainty analysis

The MECAN software package [12] has been used for preprocessing of the individual data, regression, calculation of risk estimates, probabilities for radiation-induced cases and simulation of uncertainty intervals. MECAN includes the C++ library Minuit2 [13] from CERN which is used for the minimization of $-2\ln\mathcal{L}$ where $\mathcal{L}$ denotes the individual likelihood function of Eq. (S17). In contrast to Poisson regression, stratification into a person year table is not required by individual likelihood regression. The deviance is given by the minimum of $-2\ln\mathcal{L}$ which is reached with the maximum likelihood estimates of the model parameters. It is assumed that a parabolic approximation of the region is valid around the minimum. In this case Wald-based standard errors (SE) and a correlation matrix can be computed for the model parameters. Confidence intervals for risk estimates (CI) are calculated by Monte-Carlo simulation based on the Delta method [14]. Results of MECAN have been compared with the EPICURE package [15] and found to be in good agreement. More details on statistical analysis with MECAN are given in refs. [8, 16, 17].

## Utilizing molecular biomarkers for risk assessment

In our epidemiological analysis the conditional probability of finding the CLIP2 marker in a PTC patient is given by

$$P_{epi}(C2C|\text{PTC}) \text{ in } [a, a + \Delta a] = \frac{P_{epi}(C2C) \text{ in } [a, a + \Delta a]}{P_{epi}(\text{PTC}) \text{ in } [a, a + \Delta a]} = \frac{h_{C2C}(a)}{h(a)} \quad \text{(S18)}$$

for attained age $a$ in a short interval $\Delta a$. $P_{epi}(\text{PTC}) = h(a)S(a)\Delta a$ denotes the probability of occurrence for a PTC case and $P_{epi}(C2C) = h_{C2C}(a)S(a)\Delta a$ denotes that same probability for a CLIP2-positive PTC case. The total hazard $h = h_{MSC} + h_{C2C}$ of the preferred mechanistic model (Fig. 1 of the main paper) is decomposed into hazard functions for the MSC and CSC part. The CSC part produces CLIP2-positive PTCs either sporadically or induced by radiation with corresponding hazard functions $h_{C2C} = h_{sC2C} + h_{rC2C}$. Analytical expressions are given in Eqs. (S5),(S8), (S11) and (S12). $P_{epi}(C2C|\text{PTC})$ has been calculated from hazard functions which were obtained from model fits to the epidemiological UkrAm cohort without making use of the CLIP2 status explicitly. We just assumed that latent information on the CLIP2 status facilitates the identification of model parameters. [1]

In molecular biology the probability $P_{mol}$ of finding a CLIP2 marker in a PTC case has been derived from logistic regression on molecular measurements [18]. For young patients (AaO < 20 yr) $P_{mol}$ shows a strong dose response (Fig. 2 of main paper or Fig. 2 of ref. [18]).

To create a link between molecular biology and radiation epidemiology, $P_{mol}$ is compared with the probability of causation

$$PC = P_{epi}(rC2C|\text{PTC}) = \frac{h_{rC2C}}{h} = \frac{\text{ERR}}{1 + \text{ERR}}. \quad \text{(S19)}$$

$PC$ quantifies the contribution of radiation to the occurrence of a tumor from statistical association [19]. It is determined by the Excess Relative Risk ERR $= h/h_0 - 1$ with $h_0 = h_{MSC} + h_{sC2C}$ as baseline hazard. To establish CLIP2 as a radiation marker, the baseline hazard $h_0$ must not include CLIP2-positive PTCs in the conceptual model. This is true for AaO < 20 yr where $h_{sC2C} \simeq 0$ so that $P_{epi}(rC2C|\text{PTC}) \simeq P_{epi}(C2C|\text{PTC})$. Good agreement of the dose response curves for $P_{mol}$ and $P_{epi}$ as shown in main Fig. 2 validates the role of CLIP2 as radiation biomarker.

---

[1]Note, that a similar assumption holds for descriptive risk models which determine the number of radiation-induced cases among the total cases in a radio-epidemiological cohort.

From our present epidemiological analysis it is impossible to determine the precise CLIP2 status of a patient $i$ with the mechanistic model. But a probability

$$
\begin{aligned}
P_{epi,i}(C2C|\text{PTC}) &= P_{epi,i}(rC2C|\text{PTC}) + P_{epi,i}(sC2C|\text{PTC}) \quad \text{(S20)} \\
&= \frac{h_{rC2C,i}}{h_i} + \frac{h_{sC2C,i}}{h_i}
\end{aligned}
$$

can be assigned individually to estimate the number of CLIP2-positive PTCs $\hat{N}(x)$ in a given group of $N_g$ patients by using

$$
\hat{N}(x) = \sum_{i=1}^{N_g} P_{epi,i}(x|\text{PTC}) \quad \text{with} \quad x = \left\{ \begin{array}{l} rC2C : \text{radiation-induced} \\ sC2C : \text{sporadic} \\ C2C : \text{all CLIP2+.} \end{array} \right. \quad \text{(S21)}
$$

$N_g$ denotes the number of patients in cohorts Genrisk-T, Genrisk-T plus and UkrAm (CTB tumors only) in Table 2 of the main paper or the number of patients in different dose and age groups in Table S7. Comparison of model estimates for $\hat{N}(C2C)$ to the measured numbers $N_{C2C}$ in different patient groups is applied for model validation.

# Results

## Preferred risk models

Table S4: Deviance and number of adjusted model parameters $N_{par}$ for selected descriptive and mechanistic models.

| model type | dose response | deviance | $N_{par}$ |
|---|---|---|---|
| descriptive | | | |
| ERR | linear, unisex | 1700.7 | 4 |
| EAR | linear, unisex | 1681.7 | 4 |
| EAR | linear, sex-spec. | 1671.7 | 5 |
| [a]EAR | linear-exp., sex-spec. | 1665.8 | 6 |
| mechanistic, two paths MSC and C2C | | | |
| rC2C only | rate $\mu_0$, linear, unisex | 1683.1 | 4 |
| rC2C only | rate $\mu_0$, linear, sex-spec. | 1672.8 | 5 |
| rC2C only | rate $\mu_0$, linear-exp., sex-spec. | 1667.2 | 6 |
| [b]rC2C & sC2C | rate $\mu_0$, linear-exp., sex-spec. | 1665.8 | 6 |

[a]preferred descriptive model, Eq. (S13)
[b]preferred mechanistic model, Eq. (S12), sporadic CLIP2-associated carcinogenesis (sC2C) is described by Eq. (S8) with 3 additional model parameters fixed

Table S5: Parameters of the preferred mechanistic model, maximum likelihood estimates (MLE) with Wald-based 95% confidence intervals (CI) in brackets.

| symbol | unit | MLE (95% CI) | relation to biological parameters |
|---|---|---|---|
| MSC: sporadic path, CLIP2 biomarker negative | | | |
| $R_0$ | $yr^{-3}$ | -15.7 (2.5)[a] | $= N_s \nu_0 \nu_1 \nu_2$ |
| $\gamma$ | $yr^{-1}$ | 0.11 (0.15) | $= \alpha - \beta$ |
| sC2C: sporadic path, CLIP2 biomarker positive | | | |
| $\hat{X}_0$ | $yr^{-3}$ | -11.93 (fixed)[a] | $= N_b \mu_0 \hat{\mu}_1$ |
| $s_{\mu_1}$ | $yr^{-1}$ | 0.5 (fixed) | |
| $a_{\mu_1}$ | yr | 22 (fixed) | |
| rC2C: radiation-induced path, CLIP2 biomarker positive | | | |
| $X_0$ | $yr^{-3}$ | -23.96 (fixed)[a] | $= N_b \mu_0 \mu_1$ |
| $\kappa$ | $Gy^{-1}$ | 20.58 (0.42)[a] | |
| $\varepsilon$ | $Gy^{-1}$ | -0.0894 (0.87) | |
| $\sigma$ | - | 0.43 (0.29) | |
| $\Phi_{obl86}$ | - | -0.92 (0.46) | |

[a]log-transformed

Table S6: Parameters of the preferred descriptive model, maximum likelihood estimates (MLE) with Wald-based 95% confidence intervals (CI) in brackets.

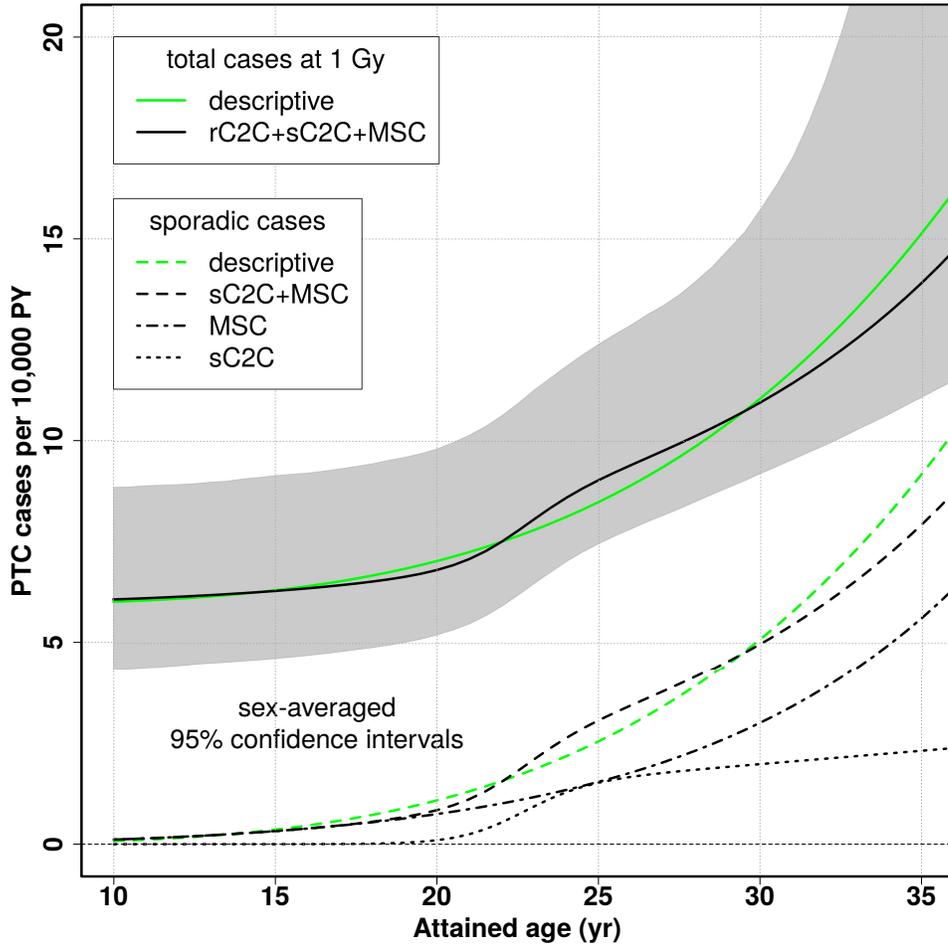| Symbol | Unit | MLE (95% CI) |
|---|---|---|
| baseline $h_0$ | | |
| $b_0$ | - | 0.94 (0.46) |
| $b_a$ | - | 3.8 (1.8) |
| dose response $EAR(s, D)$ | | |
| ear | $10^{-4} PY^{-1} \, Gy^{-1}$ | 6.5 (2.7) |
| $b_{exp}$ | $Gy^{-1}$ | -0.0891 (0.87) |
| $b_s$ | - | 0.43 (0.29) |
| $\Phi_{obl86}$ | - | -0.93 (0.46) |

Figure S3: Decomposition of the sex-averaged total hazard function (PTC cases per 10,000 PY) according to Eq. (S12) for the preferred mechanistic model (black lines) into components for sporadic CLIP2-associated carcinogenesis (sC2C, Eq. (S8)), sporadic multi-stage carcinogenesis (MSC, Eq. (S11)) and radiation-induced CLIP2-associated carcinogenesis (rC2C, Eq. (S5)), gray-shaded 95% CI from the mechanistic model; for comparison baseline and total hazard of the preferred descriptive model (green lines) are shown.
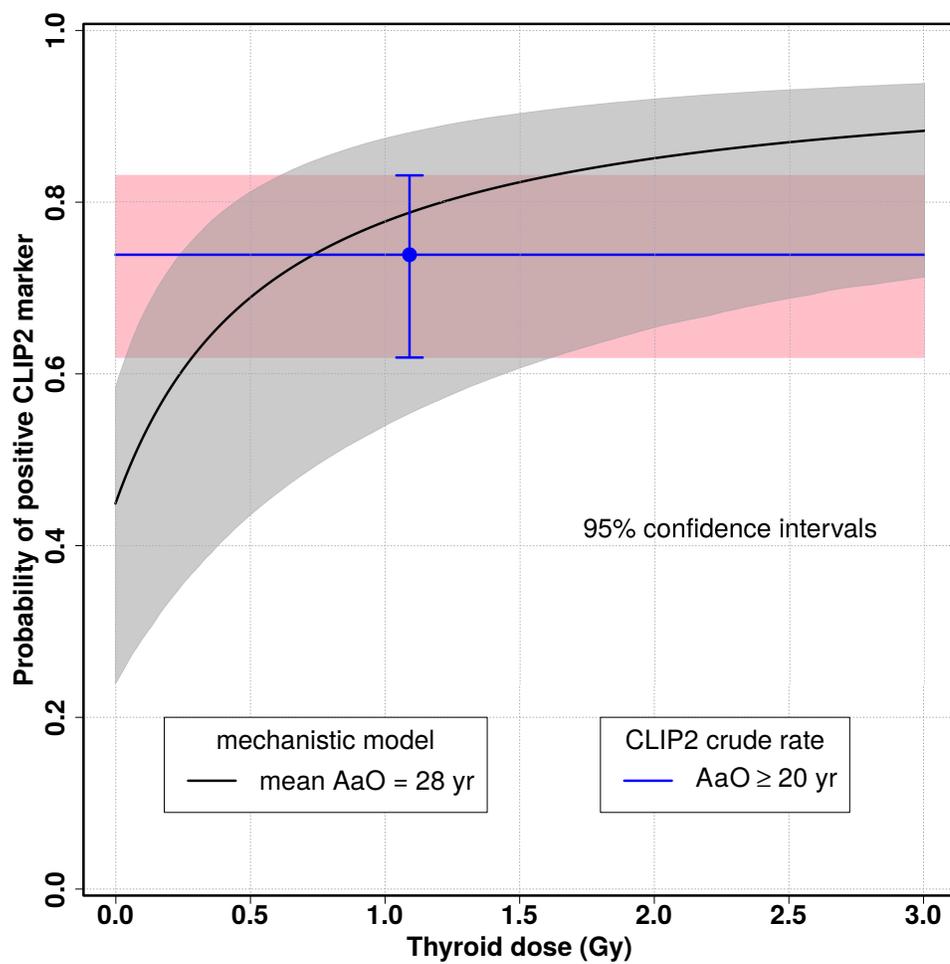
## Validation with the CLIP2 marker



Figure S4: Dose response for $P_{epi}(C2C|\text{PTC})$ at attained age 28 yr (mean age in group AaO $\geq$ 20 yr) from mechanistic model (black line) compared to the measured crude rate of the positive CLIP2 marker among PTCs with AaO $\geq$ 20 yr [18] (blue point) from Table S7.

Table S7: Occurrence of CLIP2 marker in 141 PTCs used in Selmansberger et al. [18] for old patients (AaO $\geq$ 20 yr) and for young patients (AaO < 20 yr) in dose groups < 0.2 Gy, $\geq$ 0.2 Gy - < 1 Gy and $\geq$ 1 Gy; crude rate $N_{C2C}/N_g$ of PTCs with positive CLIP2 marker from measurements, and predicted share of PTCs $\hat{N}(C2C)/N_g$ with positive CLIP2 marker from mechanistic model using Eq. (S21).

| | all doses | | |
|---|---|---|---|
| AaO $\geq$ 20 yr | arithmetic mean (standard deviation) | | |
| AaE (yr) | 8.9 (4.1) | | |
| TsE (yr) | 17.8 (2.8) | | |
| AaO (yr) | 26.7 (4.1) | | |
| thyroid dose (Gy) | 1.09 (2.17) | | |
| cases $N_{C2C}/$ in group $N_g$ | 48/65 | | |
| $N_{C2C}/N_g$, crude rate (95% CI) | 0.74 (0.63; 0.84) | | |
| $\hat{N}(C2C)/N_g$, model | 0.66 (43.3 cases) | | |
| dose group | < 0.2 Gy | $\geq$0.2 Gy - < 1 Gy | $\geq$ 1 Gy |
| AaO < 20 yr | arithmetic mean (standard deviation) | | |
| AaE (yr) | 2.1$^a$ (1.2) | 1.6 (1.4) | 2.2 (1.9) |
| TsE (yr) | 15.8$^a$ (1.5) | 15.4 (1.5) | 15.2 (1.5) |
| AaO (yr) | 16.2 (2.7) | 16.9 (1.8) | 17.4 (1.6) |
| thyroid dose (Gy) | 0.057 (0.058) | 0.461 (0.223) | 2.43 (1.40) |
| cases $N_{C2C}/$ in group $N_g$ | 18/41 | 17/23 | 10/12 |
| $N_{C2C}/N_g$, crude rate (95% CI) | 0.44 (0.29; 0.59) | 0.74 (0.54; 0.91) | 0.83 (0.58; 1) |
| $\hat{N}(C2C)/N_g$, model | 0.42 (17.2 cases) | 0.85 (19.6 cases) | 0.96 (11.5 cases) |

$^a$for exposed cases, 24 patients with nominal doses of 1 mGy/yr were born after 1986.

# Risk assessment

Table S8: Maximum likelihood estimates (95% CI) from mechanistic model (upper table) and from descriptive model (lower table) for the EAR (PTC cases per 10,000 PY at 1 Gy) and for the ERR at 1 Gy and attained age 16, 24, 27 and 35 yr; case-weighted averaging over all oblasts reduces the EAR by a factor of 0.82 compared to Kyiv and Chernihiv oblast.

| mechanistic model | male | female | both sexes |
|---|---|---|---|
| | EAR (PTC cases per 10,000 PY at 1 Gy) | | |
| Kyiv, Chernihiv ob. | 3.9 (2.2; 6.9) | 9.2 (6.2; 13) | 6.0 (4.1; 8.8) |
| all oblast-average | 3.3 (1.9; 5.9) | 7.8 (5.3; 11) | 5.1 (3.5; 7.5) |
| attained age (yr) | ERR at 1 Gy | | |
| $16^a$ | 9.3 (1.4; 45) | 22 (3.8; 91) | 14 (2.4; 62) |
| $24^b$ | 1.4 (0.49; 2.8) | 3.2 (1.3; 5.5) | 2.1 (0.82; 3.7) |
| $27^c$ | 1.0 (0.39; 2.0) | 2.4 (1.1; 4.0) | 1.6 (0.67; 2.6) |
| 35 | 0.48 (0.16; 0.94) | 1.2 (0.39; 2.0) | 0.75 (0.26; 1.3) |

| descriptive model | male | female | both sexes |
|---|---|---|---|
| | EAR (PTC cases per 10,000 PY at 1 Gy) | | |
| Kyiv, Chernihiv ob. | 3.9 (2.0; 6.4) | 9.1 (5.8; 13) | 6.0 (3.6; 8.2) |
| all oblast-average | 3.3 (1.7; 5.5) | 7.8 (5.0; 11) | 5.1 (3.1; 6.7) |
| attained age (yr) | ERR at 1 Gy | | |
| $16^a$ | 7.6 (1.8; 29) | 18 (4.8; 59) | 12 (2.9; 40) |
| $24^b$ | 1.7 (0.64; 3.8) | 3.9 (1.8; 7.5) | 2.6 (1.1; 5.1) |
| $27^c$ | 1.1 (0.47; 2.4) | 2.7 (1.4; 4.7) | 1.7 (0.83; 3.1) |
| 35 | 0.42 (0.18; 0.85) | 0.99 (0.51; 1.8) | 0.64 (0.32; 1.2) |

[a]mean age at operation in Tronko et al. [1], their ERR 5.3 (1.7; 28) for all histological types in $1^{st}$ screening
[b]mean age at operation in present study
[c]mean age at operation in Brenner et al. [2], their ERR 1.6 (0.30; 5.4) for PTC only in $2^{nd}$ to $4^{th}$ screening

## Analysis of molecular measurements

Table S9: Estimated cluster-specific share of MSC-associated cases and radiation-induced cases from mechanistic model and mean values for age at exposure (AaE), time since exposure (TsE), age at operation (AaO) and thyroid dose in the four clusters of copy number alterations (CNAs) from Selmansberger et al. [20] (their Figure 2) for 79 PTC cases with TDose10 estimates.

| cluster | CNAs | cases | MSC cases | MSC share | radiat.-induced cases | radiat.-induced share | AaE (yr) | TsE (yr) | AaO (yr) | dose (Gy) |
|---------|------|-------|-----------|-----------|-----------------------|-----------------------|----------|----------|----------|-----------|
| 1.1 | few[a] | 16 | 4.6 | 0.29 | 8.7 | 0.54 | 8.7 | 17.1 | 25.8 | 1.11 |
| 1.2 | many[b] | 14 | 5.4 | 0.39 | 5.3 | 0.38 | 8.1 | 20.0 | 26.0 | 0.326 |
| 2.1 | few[a] | 41 | 10.9 | 0.27 | 23.2 | 0.57 | 7.6 | 16.7 | 24.3 | 1.66 |
| 2.2 | many[b] | 8 | 2.8 | 0.35 | 3.3 | 0.42 | 10.0 | 16.2 | 26.2 | 0.643 |
| total | | 79 | 23.7 | 0.30 | 40.5 | 0.51 | 8.2 | 17.3 | 25.5 | 1.22 |

[a]most chromosomes carry few ($\leq 3$) CNAs, or are 'CNA-silent'
[b]most chromosomes carry CNAs distributed at random, except 1p, 19, 22 with cumulative occurrence

Table S10: Odds ratios of radiation-induced case shares for PTC cases with BRAF mutations, RET/PTC rearrangements and with transcriptomic profiles in four clusters 'RAS-like' (C1), 'intermediate RAS-BRAF' (C2), 'BRAF-like' (C3) and radiation-related (C4) from Selmansberger et al. [20], their online Fig. S1, number of radiation-induced PTC cases predicted by mechanistic model, due to low case numbers the odds ratios represent only trends.

| molecular | pro- | PTC cases | | pro- | PTC cases | | odds |
| alterations | perty | measured | rad.-ind. | perty | measured | rad.-ind. | ratio |
|---|---|---|---|---|---|---|---|
| BRAF mutations | positive | 8 | 3.0 | negative | 57 | 32.2 | 0.65 |
| RET/PTC rearrangemts. | positive | 19 | 11.4 | negative | 44 | 21.9 | 1.2 |
| transcriptomic cluster | C4 | 9 | 6.1 | C1, C2, C3 | 13 | 5.5 | 1.6 |

# Discussion

Table S11: Path-specific case shares and corresponding odds ratios (ORs) from mechanistic model for epidemiological covariables of age at exposure (AaE), time since exposure (TsE), age at operation (AaO) and thyroid dose in radiation-induced (rC2C) and sporadic (sC2C) CLIP2-associated carcinogenesis and in multi-stage carcinogenesis (MSC), tumor cohort from Selmansberger et al. [18].

| epidem. covar. | condi- tion | PTC cases yes | no | rC2C share yes | no | OR (yes/no) | sC2C share yes | no | OR (no/yes) | MSC share yes | no | OR (no/yes) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AaE[a] | < 5 yr | 63 | 54 | 0.75 | 0.41 | 1.8 | 0.03 | 0.24 | 8.1 | 0.22 | 0.35 | 1.6 |
| TsE[a] | < 17 yr | 73 | 44 | 0.67 | 0.48 | 1.4 | 0.10 | 0.18 | 1.8 | 0.23 | 0.35 | 1.5 |
| AaO[b] | < 20 yr | 76 | 65 | 0.63 | 0.44 | 1.4 | 0.01 | 0.22 | 30 | 0.36 | 0.34 | 0.92 |
| dose[b] | ≥ 0.2 Gy | 67 | 74 | 0.75 | 0.31 | 2.5 | 0.08 | 0.14 | 1.8 | 0.17 | 0.55 | 3.3 |

[a]117 exposed cases born before 1986 with TDose10 estimates
[b]141 cases, including 24 cases born after 1986 with nominal doses of 1 mGy/yr

# References

[1] Tronko MD et al. (2006) A cohort study of thyroid cancer and other thyroid diseases after the chornobyl accident: thyroid cancer in Ukraine detected during first screening. *J. Natl. Cancer Inst.* 98(13):897–903.

[2] Brenner AV et al. (2011) I-131 dose response for incident thyroid cancers in Ukraine related to the Chornobyl accident. *Environ. Health Perspect.* 119(7):933–9.

[3] Little MP, Wright EG (2003) A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Math Biosci* 183(2):111–34.

[4] Little MP, Vineis P, Li G (2008) A stochastic carcinogenesis model incorporating multiple types of genomic instability fitted to colon cancer data. *Journal of Theoretical Biology* 254:229–238.

[5] Moolgavkar SH, Knudson AGJ (1981) Mutation and cancer: a model for human carcinogenesis. *J Natl Cancer Inst* 66(6):1037–52.

[6] Feller W (1968) *An Introduction to Probability Theory and Its Applications*, Wiley Series in Probability and Mathematical Statistics. (Wiley, New York) Vol. 1, 3rd edition.

[7] Heidenreich WF, Hoogenveen R (2001) Limits of applicability for the deterministic approximation of the two-step clonal expansion model. *Risk Anal.* 21(1):103–5.

[8] Kaiser JC, Meckbach R, Jacob P (2014) Genomic instability and radiation risk in molecular pathways to colon cancer. *PLoS ONE* 9(10):e111024.

[9] Armitage P, Doll R (1957) The Two-Stage Theory of Carcinogenesis in Relation to the Age Distribution of Human Cancers. *Br. J. Cancer* 11:161–169.

[10] Little MP, Heidenreich WF, Li G (2010) Parameter identifiability and redundancy: theoretical considerations. *PLoS One* 5(1):e8915.

[11] Heidenreich WF, Jacob P, Paretzke HG, Cross FT, Dagle GE (1999) Two step model for fatal and incidental lung tumor risk in rats exposed to radon. *Radiation Research* 151:209–217.

[12] Kaiser JC (2010) MECAN - A Software Package to Estimate Health Risks in Radiation Epidemiology with Multi-Model Inference, User's Guide.

[13] Moneta L, James F (2010) Minuit2 Minimization Package User's Guide, (CERN), Report. v. 5.27.02, accessed 20 February 2012.

[14] Casella G, Berger RL (2002) *Statistical Inference.* (Duxbury, Pacific Grove, CA), 2 edition.

[15] Preston DL, Lubin JH, Pierce DA (1993) Epicure User's Guide.

[16] Kaiser JC, Jacob P, Meckbach R, Cullings HM (2012) Breast cancer risk in atomic bomb survivors from multi-model inference with incidence data 1958-1998. *Radiation and Environmental Biophysics* 51(1):1–14.

[17] Kaiser JC, Walsh L (2013) Independent analysis of the radiation risk for leukaemia in children and adults with mortality data (1950-2003) of Japanese A-bomb survivors. *Radiation and Environmental Biophysics* 52(1):17–27.

[18] Selmansberger M et al. (2015) Dose-dependent expression of CLIP2 in post-Chernobyl papillary thyroid carcinomas. *Carcinogenesis* 36(7):748–56.

[19] Kocher DC et al. (2008) Interactive RadioEpidemiological Program (IREP): a web-based tool for estimating probability of causation/assigned share of radiogenic cancers. *Health Phys* 95(1):119–47.

[20] Selmansberger M et al. (2015) Genomic copy number analysis of Chernobyl papillary thyroid carcinoma in the Ukrainian-American Cohort. *Carcinogenesis* 36(11):1381–7.