UDC 57.042; 519.246.2
DOI: 10.21668/health.risk/2021.2.02.eng

Read online

Research article

GROUP HEALTH RISK PARAMETERS IN A HETEROGENEOUS COHORT. INDIRECT ASSESSMENT AS PER EVENTS TAKEN IN DYNAMICS

V.F. Obesnyuk

The Southern Urals Biophysics Institute of the RF Federal Medical and Biological Agency, 19 Ozerskoe drive, Ozersk, 456780, Russian Federation

The present work focuses on describing a procedure for assessing intensive and cumulative parameters of specific risk when observing cohorts under combined exposure to several external or internal factors.

The research goal was to reveal how to use well-known heuristic-descriptive parameters accepted in remote consequences epidemiology for analyzing dynamics of countable events in a cohort; analysis should be performed on quite strict statistic-probabilistic grounds based on Bayesian approach to explaining conditional probabilities that such countable events might occur. The work doesn't contain any new or previously unknown epidemiologic concept or parameters; despite that, it is not a simple literature review. It is the suggested procedure itself that is comparatively new as it combines techniques used to process conventional epidemiologic information and a correct metrological approach based on process description.

The basic result is providing a reader with understanding that all basic descriptive epidemiologic parameters within cohort description framework turn out to be quantitatively interlinked in case they are considered as conditional group processes. It allows simultaneous inter-consistent assessment of annual risk parameters and Kaplan – Meier (Fleming – Harrington) and Nelson – Aalen cumulative parameters as well as other conditional risk parameters or their analogues. It is shown that when a basic descriptive characteristic of cumulative parameters is chosen as a measure for measurable long-term external exposure, it is only natural to apply such a concept as a dose of this risk factor which is surrogate in its essence. Operability of the procedure was confirmed with an example. The suggested procedure was proven to differ from its prototype that previously allowed achieving only substantially shifted estimates, up to ~100 % even in case an operation mode was normal. Application requires creating specific but quite available PC software.

Key words: *risk, parameter, epidemiology, risk factor, competition, indirect estimate, mortality, process, cohort, strata, model.*

There is a well-established opinion on a health risk being an objective probability that this or that undesirable event will occur in future [1–3] due to certain conditions/factors including a period of observation. Such examined undesirable events are usually death, a disease or, a bit less frequently, sub-clinical irreversible changes in a person's health that are reliably diagnosed. Accordingly, a risk can be given quantitatively as a certain numeric or functional indicator. In prediction practices it can characterize an event that hasn't yet occurred; however, it is rather difficult to reach a stage at which risks can be managed if experience gained via observing similar events in similar conditions hasn't been generalized, that is, risks have not been assessed a posteriori as per previously collected data. In that respect a task related to risk measuring is similar to a metrological procedure used for determining a certain unknown (but objectively existing) value or to an establishing a correlation between a risk indicator and conditions for its potential realization. This logical scheme contains a seeming internal contradiction related to the fact that a concept of probability implies "randomness" category being active and this category contradicts "link" category since the latter is a determined one. Yes, randomness is present here and it produces its effects but still there is no contradiction. It makes sense to apply "risk" category only if there is a probable

Valery F. Obesnyuk – Candidate of Physical-Mathematical Sciences, Associate Professor, Senior researcher (e-mail: v-f-o@subi.su; tel.: +7 (35130) 7-52-36; ORCID: https://orcid.org/0000-0002-2446-4390).

alternative course of events. However, risk factor can be present here as this or that determined combination thus creating "factors – risk" link which is quite real; this dependence should be examined thoroughly prior to a stage at which risks are managed.

Let us note that "risk" as a concept has certain properties that are purely mathematical and make risk assessment procedures rather complicated. A link or a potential link with risk factors indicates that we deal with a conditional probability. Further complications arise due to time or age being usually taken as a risk factor when specific health risks are analyzed; therefore, we can conclude that health risk is not only an indicator (a number) but also a dynamic random process. And finally, if there is a task aimed at eliminating influences exerted on a metrological procedure by random or uncontrollable factors, this risk can never be assessed individually since assessment is possible only for a homogenous group of individuals as a certain biological property which is common for them.

Intensive risk rate[1] of common or specific mortality or morbidity, which is also known as "force of mortality", "hazard rate", or "instantaneous incidence rate", is a typical and widely spread rate used in descriptive occupational epidemiology, clinical epidemiology, medical-ecological and demographic research [4, 5]. When remote consequences are described using this value, it is usually attributed to a year on an age scale or calendar scale as the most commonly used time unit. A series of risk rates is usually used as a measurement due to its dynamics being quite reproducible when describing remote consequences of a wide range of effects, for example, non-communicable diseases for a great number of isolated sub-cohorts or sub-populations that live in similar socioeconomic conditions. It allows considering an intensive risk rate taken in its overall dynamics practically as a species-specific property. This circumstance, for example, is a reason for regular regional screening of all oncologic morbidity and mortality exactly as per the above mentioned rate [4]. As a rule, it is exactly this parameter with its excessive values being equal to 0.001–1 ‰ per year serving as permissible risk limits[2] which is taken by regulatory authorities as a sign that it is time to make relevant decisions. Intensive group risk rate is also known as "individual risk", however, this name is incorrect since it contradicts its group essence.

An insight into the given rate and its direct link with risk value and other objective parameters can be easily derived from a simple example published in the work [5] and given in Table 1.

Table 1

An example of a 5-year cohort study

	Exposed	No exposed
Died due to the cause	30	10
Didn't die due to the cause	70	90
Total	100	100

In this example, two population groups (strata) with practically the same structure were observed over a relatively short period of time $T = 5$; the only difference between them was that one group was exposed to a certain risk factor while the other was not, and it is influence exerted by this factor that we are trying to assess. If a risk is a probability that a person will die due to the examined cause, than its assessment amounts to obvious 30 out of 100 cases in the exposed group or $R_e = 0.30$; similarly, in the non-exposed group $R_n = 10/100 = 0.10$. Then excess risk of death due to the examined cause amounts to $0.3–0.1 = 0.2$; it is quite natural to relate it to effects produced by exposure to a risk factor. Relative risk for these effects is calculated as $RR = R_e / R_n = 3.0$.

These given values are cumulative death cases over the observed 5-year period. They give

[1] Epidemiologic glossary. In: D.M. Last for the International epidemiologic association, eds. 4th edition, 316 p.

[2] R 2.1.10.1920-04. The guide on assessing population health risks caused by exposure to chemicals that pollute the environment. Moscow, The Federal Center for State Sanitary and Epidemiologic Surveillance of the RF Public Healthcare Ministry Publ, 2004, 143 p.

an opportunity to assess intensive rates relaying on the following ratios $M_e = N_e \cdot \left(1 - \exp\left(-h_e T\right)\right)$ and $M_n = N_n \cdot \left(1 - \exp\left(-h_n T\right)\right)$ as well as on an assumption that intensive rates are constant in both exposed and non-exposed group. Here M_e, M_n are a number of "cases" in the exposed and non-exposed strata; N_e, N_n is an initial number of people in the strata; h_e, h_n are "hazards" or annual risk rates. Exponents occurring in these formulas indicate that h_e, h_n values are sliding, that is, related to a condition that a person reaches the current age within the observed period; whereas cumulative rates R_e, R_n refer to the whole observation period in comparison with the initial state of sub-strata. Due to it, h_e, h_n values are similar to continuous rate of discounting in economic theory as per their mathematical properties thus creating a link with exponential ratios. Their calculation brings the following results: $h_e \approx 71$ ‰ per year and $h_e \approx 21$ ‰ per year accordingly. Hazard ratio is $HR = h_e / h_n = 3.38 \neq RR = 3.0$ under effects produced by a factor.

If we neglect this ambiguous interpretation of a relative risk and return to the heading of our work, it is quite relevant to ask – why is it a problem to assess rates in a heterogeneous cohort? It seems so simple if we look at the example given in Table 1. But at the same time, methodological issues here are multiple.

1) $h = h(t)$ rate is not actually a number but a function of an age or time, that is, a process; whereas assessments given in Table 1 were reduced to simple scalar (numeric) values;

2) Table 1 is based on only one factor that influences the risk whereas when it comes down to a real sampling or a cohort, it is almost always a multi-factor study. It requires a specific procedure for statistical assessing that involves stratifying a heterogeneous cohort into more than two strata taking into account all relevant combinations of risk factors;

3) as we can see, neither h intensity nor its cumulative analogue $h \cdot T$ are not directly observed values. Contradicting (and outdated) beliefs are still alive due to a known approximated property of the value h that allows calculating it as a "ratio of a number of specific cases to a number of person-years under risk"[1]. But it is a mistake to believe that this approximated property is an exact definition. In fact, it is dynamics of countings given in Table 1, both in cumulative and individual forms, that is initial empirically observed data. This circumstance leads to a task to accomplish an ***indirect assessment*** of $h(t)$ process or its cumulative analogue as per observing countings accumulation in each homogenous stratum in the cohort;

4) countings in a homogenous stratum which is a part of a random sampling, are also random. It is necessary to assess a parameter of a certain homogenous and almost general aggregate. Due to it parameters can be assessed with certain relative uncertainty which tends to be the greater, the fewer is a number of cases in examined cohort/strata. For example, interval assessments of cumulative risks for the examined exposed and non-exposed strata (with 95 % confidence probability) amount to $R_e = 0.219 \ldots 0.396$ and $R_n = 0.056 \ldots 0.175$, and we can clearly see that uncertainty range is wider than the central assessment for the non-exposed group. It practically forbids us to work with small strata with a number of "cases" in them being lower than 4 since extended relative risk uncertainty will certainly exceed 100 %. It is completely impossible to reliably identify any process relaying on fewer than 4 points although statistically significant differences between strata may occur even with a smaller number of cases [6]. Therefore, it is necessary to create such an assessment algorithm that could preserve all advantages gained due to factor space stratification together with an opportunity to indentify ***optimal model dependence*** for risks that takes into account relations with all risk factors for the whole set of strata simultaneously.

Therefore, it is vital and natural to make an attempt to create an algorithm for assessing intensive and cumulative specific risk rates in a heterogeneous cohort basing on data obtained via a long-term observation period.

Description of the assessment procedure and its prototype. Let us note that an issue related to heterogeneity of actual observations has long been of interest when it comes down both to cohort samplings and population studies on medical and demographic problems [7]. There are a lot of established various reasons for heterogeneity that are observed while sampling representatives are still alive or revealed after their death including non-observable hidden factors [8].

Without claiming to cover everything, we are going to concentrate on examining influence exerted by only a priori known risk factors and to assume that latent variables are absent. We can only rely on a researcher-physician's intuition when he or she keeps registers and collects initial epidemiologic data. It allows grouping individual by strata even before mathematical data analysis starts with the possibility to permanently bind them to their predefined strata during the entire observation period. It is implicitly assumed that all people in a specific stratum have the same chance to fall sick or die to any examined disease coded in the ICD-9 or ICD-10. It is especially easily achieved regarding risk factors that can be described via binary attributes, for example, sex, or smoking status (smokes / doesn't smoke). Even an interfering disease in case history in a period of observation can be a binary attribute. Such factors can be considered almost immutable over a long period of time. Certain quantitative factors that exert their impacts on health can also turn out to be useful for the chosen stratification scheme in case intensity of their influence is the same for all cohort members. It is obvious for acute single exposure or chronic even exposure with the same individual intensity. In this case, it is possible to introduce such a (surrogate) cumulative factor as a dose accumulated by the end of observation period. In this case it seems only natural to analyze conditional dependence for "cumulative risk – cumulative dose" pair.

This approach has its functional prototype in epidemiology, AMFIT module in Epicure software package[3]. This software has been successfully applied in epidemiologic research all over the world [9, 10]. In particular, it was approved on as the software standard for radiation and epidemiologic studies: "… Epicure is the de-facto standard for modeling radiation health effects …" [11]. Practically all national standards for radiation safety in countries where radiation-hazardous objects are located, including Russia[4], are based on results obtained with AMFIT (Poisson regression). And it is considered to be established that an equivalent radiation exposure dose is a commonly recognized risk factor of remote radiation-oncologic consequences.

The approach which we suggest in this work is a bit different from AMFIT in spite of common goals. There are three basic differences:

a) if we stick to strictly probabilistic approach, then observations can't be described and composed function of their aggregate assessment can't be built via applying Poisson distribution for countings of "cases" in strata since Poisson statistics is suitable only for describing rare events in an unlimited sampling. Actual cohorts cease to be unlimited even in the rough sooner or later as their members die. It is necessary to describe deviations from a basic trend (process) more correctly basing on binomial statistics which is suitable for limited samplings and is not bound to a condition that events should be rare;

b) AMFIT algorithm is based on assessing intensity of specific events at the expense of assessing related cumulative parameters. Meanwhile, intensity is not a directly observed descriptive characteristic; however, the Kaplan – Meier cumulative estimator can be successfully used instead of it since it is directly linked to individual countings of "cases" and random binomial deviations. Application of this rate would allow achieving more sustainable

[3] EPICURE User's Guide. In: D. Preston, J. Lubin, D. Pierce, M. McConney eds. Seattle, Hirosoft, 1988–1993, 334 p.

[4] SER 2.6.1.2523-09. Radiation safety standards (NRB-99/2009). Moscow, The Federal Center for Hygiene and Epidemiology of Rospotrebnadzor, 2009, 100 p.

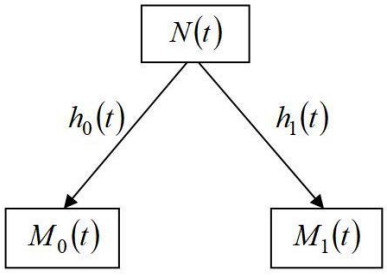

Figure 1. Markov structure scheme for dropping out from observation for two competing causes of death. There is an initial state compartment and two compartments for two registered causes of death

assessments and simultaneously preserving an opportunity to calculate intensity of events;

c) AMFIT algorithm is based on maximum likelihood as it was interpreted by Ronald Fischer which, strictly speaking, is not probabilistic and has purely heuristic basis just as its prototype, Karl Gauss' maximum likelihood [12].

Although any algorithm used for statistical processing yields biased estimates, we can still hope that measuring precision can be improved significantly due to eliminating drawbacks of AMFIT algorithm which has already been applied successfully and therefore can easily be taken as a prototype.

A relation between intensive and cumulative conditional statistical parameters and epidemiology of long-term effects. A desire to assess epidemiologic rates objectively requires applying a bit more profound mathematical apparatus than that used to process data given in Table 1. However, this mathematics shouldn't mislead anybody regarding an intention to move to analytical epidemiology sphere. Our described procedure still corresponds to common descriptive statistics which is not related to an essence of cause-and-effect relations when health is described should we refer to events in a conditionally homogenous stratum included into a heterogeneous cohort. A descriptive approach involves obvious formalism: in cases when a simple Markov scheme can be used to describe a flow of events with specific causes of deaths or diseases in a homogenous group, and this Markov scheme has three states and two competing reasons for a person to be dropped out from observation, a speed of change in a number of people in a compartment responsible for the initial state is proportionate to a number of people in it (Figure 1, Formula (1)). Regardless of how such a model is close to reality, a coefficient of proportionality between a speed of dropping out and a number of people in a basic state can also be calculated and given as a sum of two intensities of events, under study and competing.

This scheme given in Figure 1 can be roughly described with a system of common differential equations

$$\begin{aligned} \frac{dN}{dt} &\approx -\left(h_0(t)+h_1(t)\right)\cdot N, \\ \frac{dM_1}{dt} &\approx h_1(t)\cdot N, \end{aligned} \qquad (1, 2)$$

where t is time (age); $h_0(t), h_1(t)$ are intensities of events related to the examined and competing causes of death; $N(t)$ is a number of people in the initial state; $M_1(t)$ is cumulative (accumulated) number of deaths due to the examined cause. Equations (1, 2) are only approximate due to inability to calculate any derivatives from discrete-valued functions that are also susceptible to random fluctuations. However, it doesn't prevent us from moving to expected conditional fractions calculated from an initial number of people in a homogenous sub-group at a certain initial observation point t_0, and hence to conditional probabilities or prevalence in a stratum. In this case our ratios can become continuous and precise (3, 4):

$$\begin{aligned} \frac{dS}{dt} &= -\left(h_0(t)+h_1(t)\right)\cdot S, \\ \frac{dR}{dt} &= h_1(t)\cdot S, \end{aligned} \qquad (3, 4)$$

if we conditionally take $S(t_0)=1$ at a point where observation starts. Due to its linearity the system (3, 4) has a simple analytical solution:

$$S(t)=S(t|t_0)=S(t_0)\cdot P_0(t|t_0)\cdot P_1(t|t_0), \quad (5)$$

$$\Delta R\left(t|t_0\right)=\int_{t_0}^{t} h_1\left(\tau\right)\cdot S\left(\tau\right)d\tau\,, \qquad (6)$$

where $P_0\left(t|t_0\right)=\exp\left(-H_0\left(t|t_0\right)\right)$; $P_1\left(t|t_0\right)=$ $=\exp\left(-H_1\left(t|t_0\right)\right)$; $H\left(t|t_0\right)=\int_{t_0}^{t} h\left(\tau\right)d\tau$; ΔR is a growth in a risk of death due to the examined cause over a period $\left[t_0,t\right]$. As it logically comes from solutions to (5, 6) and according to experience gained by a wide circles of epidemiologists [13–23] these expressions provide an opportunity not only to know intensity of examined specific events $h_1\left(t\right)$ but also an exact survival function $S\left(t|t_0\right)$, additional conditional lifetime risk of death due to the examined cause $\Delta R\left(t|t_0\right)$, cumulative probability that a person will not die to the examined cause provided that he or she reaches an age t_0 and conditional absence of any competing causes of death – $P_1\left(t|t_0\right)$, as well as an analogue of Nelson – Aalen estimator $H_1\left(t|\,t_0\right)$ which is also known as cumulative intensity of specific mortality provided there are no other causes of death (cumulative hazard) [19, 20]. It may seem that interrelated values $P_1\left(t|t_0\right)$ и $H_1\left(t|t_0\right)$ are non-observable; however, it is not true. Epidemiologic applications of martingales theory [19] stipulate that a conditional but still quite measurable Kaplan – Meier survival function [22] corresponds to the value $P_1\left(t|t_0\right)$; and measurable and already mentioned martingale Nelson – Aalen estimator corresponds to the value $H_1\left(t|t_0\right)$. Recall that the value $H_1\left(t|t_0\right)$ which in its essence is an area below the curve showing the process $h_1\left(t\right)$ has been successfully and efficiently controlled for more than 2 decades within oncologic monitoring [4] in the Russian Federation. This parameter is convenient not only for measurements within a cohort, but also within a population.

Therefore, 6 descriptive epidemiologic parameters, 4 cumulative and 2 intensive, turn out to be related within a simple dynamic scheme shown in Figure 1. It should be noted that values of relative parameters are bound to different bases. For example, a conditional lifetime risk is calculated against a point where observation starts; but intensive parameters and their cumulative attributes $P_1\left(t|t_0\right)$ and $H_1\left(t|t_0\right)$ are sliding, that is, they are calculated against an achieved share of survived people since they are related to a condition that a person reaches a moment of observation. Besides, a true survival function turns out to be stronger related to competing causes of death and uncontrollable history of a stratum prior to a moment when observation starts in comparison with parameters that are responsible for the examined cause of death. Due to it the parameters $H_1\left(t|t_0\right)$ and $P_1\left(t|t_0\right)$ are interesting as such. Given the relation $P_1\left(t|t_0\right)=\exp\left(-H_1\left(t|t_0\right)\right)$ the parameter $P_1\left(t|t_0\right)$ can be viewed as a certain conditional analogue of a "survival function"; its peculiarity is that its limit value can fail to reach zero at $t\to\infty$ as opposed to a true survival function. This situation is quite possible in case the examined cause of death is not a leading one as opposed to a set of competing causes of death. It will also occur in such cases when fatal potential of the examined cause of death is finite due to a share of people who are potentially prone to the examined diseases being also finite. These properties allow interpreting the parameters $1-\exp\left(-H_1\left(t|\,t_0\right)\right)$ and $H_1\left(t|t_0\right)$ in a way similar to a specific risk value and almost equate them numerically but only if the limit value of Nelson – Aalen estimator doesn't exceed approximately ~ 0.1. Population cumulative mortality related to a specific localization of an oncologic disease practically always meets this condition [4]. However, for example, mortality due to all diseases of the circulatory system in a population usually overlaps this limitation and in that case $H_1\left(t|\,t_0\right)$ is not a risk assessment. $H_1\left(t|\,t_0\right)$ can also reach

relatively high values among people treated in specialized clinics or departments since they are specifically admitted there for treatment. For example, the work [24] contains assessments obtained for a group risk of death caused by prostate cancer being up to $1-P_1 \approx 24\,\%$ and it corresponds to the limit value $H_1 \approx 0.27$ over a period of time exceeding 3,000 days. For comparison, cumulative risk of prostate cancer among men in Russia doesn't exceed 5.7 % over 75 years of life [4]. Risk rate $1-\exp\left(-H_1\left(t\middle| t_0\right)\right)$ is usually abbreviated as RADS in radiation epidemiology [15, 16].

Construction principles and model parameterization. Bayesian interval statistic estimates. As was shown, estimation of cumulative rates for specific countries doesn't involve considerable technical difficulties; however, any detailed stratification of observations within factor space results in a decrease in a number of "cases" in each stratum and, accordingly, to a risk assessment becoming more and more uncertain. The only way to make assessment more exact and simultaneously preserve detailed description is to apply a unified approximating mathematical model for all strata simultaneously. Thus all the observed "cases" are included into calculations and it, provided there is relevant optimization, will allow achieving more steady risk assessments. It is exactly the role that should be played by a unified model for all strata. It should be dynamic, that is, suitable for all observations distributed as per time (age). Strictly speaking, this model should correspond to an essence of correlations between factors and risk rates; however, usually it is a research object by itself, that is, there is usually no such model until analysis is completed. In this case we can rely on expected similarity in dynamics of risk realization over time for different strata $h = h\left(t\middle| \mathbf{z}, \boldsymbol{\beta}, Data\right)$ basing on already examined trends for parameters in a certain reference group. Here $\mathbf{z}$ is a risk factors vector; $\boldsymbol{\beta}$ is a relevant vector of adjustable model parameters. When examining oncologic effects, it seems advisable to use population parameters [4] and, basing on perturbation technique, to introduce a relation with risk factors and relevant parameterization into the description. For example, we can use the fact that most intensive parameters showing a risk of death due to analogue oncologic diseases are unimodal functions if they are taken in time dynamics; these functions are characterized with approximately power-law growth within a range of ages being 60–65 years with a drastic fall at an age exceeding 75 years.

If a continuous model $h\left(t\middle| \mathbf{z}, \boldsymbol{\beta}, Data\right)$ to a certain extent is adequate to examined multiplicity of discrete empirical $Data$ countings, it is an attempt to dually describe the same events either directly via countings or within a space of parameters $\boldsymbol{\beta}$. Certain continuous conditional distribution of parameters over space will correspond to natural dispersion of observations over space. It will be interesting for interval estimation of multiple suitable parametric hypotheses if estimation procedures are given a probabilistic form.

Bayes' theorem is a suitable instrument for it since it allows linking these two abovementioned types of conditional distributions:

$$\psi\left(\boldsymbol{\beta}\middle| Data\right) = \frac{L\left(Data\middle| \boldsymbol{\beta}\right)\cdot prior\left(\boldsymbol{\beta}\right)}{\int L\left(Data\middle| \boldsymbol{\beta}\right)\cdot prior\left(\boldsymbol{\beta}\right) d\boldsymbol{\beta}} \qquad (7, 8)$$

или $\psi\left(\boldsymbol{\beta}\middle| Data\right) \propto L\left(Data\middle| \boldsymbol{\beta}\right)\cdot prior\left(\boldsymbol{\beta}\right).$

Although Bayesian approach is considered to be a direct statistical competitor for a well-known maximum likelihood procedure, both approaches are organically related to each other. Here $L\left(Data\middle| \boldsymbol{\beta}\right)$ is density of observations distribution for a fixed parametric model; $\psi\left(\boldsymbol{\beta}\middle| Data\right)$ is density of model parameters distribution for collected observations; $prior\left(\boldsymbol{\beta}\right)$ is a priori distribution of parameters in a presumably relevant "hazard" model. In a sense of conditional distributions ψ is Bayes' likelihood; L is Fischer's likelihood, and an expected area of the most probable parameters lies close to a maximum likelihood point in the function L, at least in such studies where a result is unknown until experimental data have

been analyzed. The relations (7, 8) would be quite strict if the a priori distribution $prior(\boldsymbol{\beta})$ were known; due to this an opinion is valid[5] that the concept of parametrically dependent likelihood is not identical to the concept of conditional probability density [25]. However, let us speak for Bayesian approach via mentioning that each new study, and especially a single one, is characterized with almost complete absence of pre-experimental knowledge due to which the function $prior(\boldsymbol{\beta})$ certainly has a significantly greater width than c $L(Data \mid \boldsymbol{\beta})$ as a function of parameters $\boldsymbol{\beta}$ in a certain significant area. Therefore it will not be a mistake to assume there is certain non-informative a priori distribution or even $prior(\boldsymbol{\beta}) \propto 1$ in a significant area. Then formally $\psi(\boldsymbol{\beta} \mid Data) \approx L(Data \mid \boldsymbol{\beta})$ and it is exactly what Ronald Fischer used and it still didn't prevent him from rejecting Bayesian approach completely in his publications [25]. This similarity of concepts developed within Bayesian and Fischer's likelihoods justifies considering constant parameters of likelihood function $L(Data \mid \boldsymbol{\beta})$ as adjustable model variables. What is considered a constant vector in frequentist concepts by Fischer and Pearson turns out to be a continuous random variable as per Laplas / Bayes under stricter consideration.

Let us point out the main thing here: likelihood for a set of strata due to their independence is simply equal to a product obtained via multiplying likelihoods for each homogenous stratum. Therefore, let us build likelihood for a separate homogenous stratum. To do that, we introduce our grid of time moments t_i within its limits and each node in this grid is bound to a specific event on a numerical axis showing age. The point t_0 corresponds to a beginning of observations. It is rather rare, that 2 or 3 such events occur in the same node simultaneously, therefore each semi-open interval $(t_{i-1}, t_i]$ between two neighboring nodes is related to its own quantity of accumulated specific cases m_i. Usually $m_i = 1$. Overall number of accumulated cases amounts to $M_i = \sum_{j=1}^{i} m_j$ by the examined moment of t_i.

Combined likelihood of observations over the whole sequence of specific events in the *j*-th stratum with a factor vector $\mathbf{z}_j$ as a chain of sequential transitions is

$$L_j = L(M_{i\max}, M_{i\max-1}, \ldots, M_1 \mid \mathbf{z}_j, \boldsymbol{\beta}) = \\ = 1 \cdot \prod_{i=1}^{i\max} p\left(M_i \mid M_{i-1}\right), \quad (9)$$

where

$$p(M_i \mid M_{i-1}) = \\ = \frac{N_{i-1}!}{m_i!\ (N_{i-1} - m_i)!} \left(\pi_i\right)^{N_{i-1}-m_i} \left(1-\pi_i\right)^{m_i}. \quad (10)$$

Here we also introduce $\pi_i = \exp(-H_i)$ and $H_i = H(t_i \mid t_{i-1}, \mathbf{z}, \boldsymbol{\beta})$ where increases in cumulative risk intensity are integrals of model intensity function

$$H_i = H_i\left(t_i \mid t_{i-1}, \mathbf{z}, \boldsymbol{\beta}\right) = \int_{t_{i-1}}^{t_i} h\left(\tau \mid \mathbf{z}, \boldsymbol{\beta}\right) d\tau. \quad (11)$$

The likelihood (9, 10) is differential in its structure just as Fischer's conventional likelihood for independent events; however, it is not quite true. If we analyze partial likelihoods $L_i \sim \pi_i^{N_{i-1}-m_i}\left(1-\pi_i\right)^{m_i}$, we can easily note that they reach their maximum value at $\pi_i^{opt} = \left(N_{i-1} - m_i\right)/N_{i-1}$, that is, they satisfy to Kaplan – Meier procedure locally [22] at each *i*-th time step for a homogenous stratum. Therefore, using functional (9, 10) to its maximum can potentially result in interpolation of a cumulative parameter if we consider estimates $\hat{\pi}_i = \pi_i^{opt}$ to be interpolating model parameters. Naturally, the same property holds approximately in case there are fewer parameters within a considered vector $\boldsymbol{\beta}$ but with added filtrating property of likelihood as estimating functional. Therefore, this constructed

[5] Reference book on applied statistics. In: E. Lloyd, U. Lederman, eds. Moscow, Finance and statistics Publ., 1989, 510 p.

likelihood can simultaneously provide both differential and cumulative approximation (regression). Intensive rates are estimated analogically to numerical differentiation of a changing noisy function. Empirical data differentiation is a poorly grounded (incorrect) numerical operation. On the contrary, derivatives from a smoothed cumulative function are going to be more stable.

It is technically more convenient not to use drastically changing likelihood function $L\left(Data \mid \boldsymbol{\beta}\right)$ or density function $\psi\left(\boldsymbol{\beta} \mid Data\right)$ but to operate with their doubled natural logarithm that is shifted against the ultimate saturation point (interpolation). Then, instead of searches near to a maximum in the expression (9) for one stratum, we should analyze the function

$$\Omega\left(\boldsymbol{\beta} \mid Data\right) = \\ = 2 \cdot \sum_i \left[\begin{array}{l} \left(N_{i-1} - m_i\right) \ln\left(\dfrac{N_{i-1} - m_i}{N_{i-1}\,\pi_i\left(\boldsymbol{\beta}\right)}\right) + \\ + m_i \ln\left(\dfrac{m_i}{N_{i-1}\left(1 - \pi_i\left(\boldsymbol{\beta}\right)\right)}\right) \end{array} \right] \quad (12)$$

close to its minimum. Contributions (12) are to be summed for the whole set of strata that are not empty. Ultimately in this case we can speak about achieved deviation (estimation functional)

$$\Omega_{\Sigma}\left(\boldsymbol{\beta}\right) = \sum_j \Omega_j\left(\boldsymbol{\beta} \mid Data\right); \\ \Omega_j\left(\boldsymbol{\beta} \mid Data\right) \geq 0. \quad (13)$$

According to well-known concepts [26, 27] that are typical for Fischer's approach, a value at which $\Omega_{\Sigma}\left(\boldsymbol{\beta}\right)$ deviates from zero gives grounds for making judgments on quality and statistical significance of completed approximation; and models are to be selected basing on difference in achieved optimal values. If parametric deviation (13) is close to a quadratic one as per small offset from the center (that is, $\psi\left(\boldsymbol{\beta} \mid Data\right)$ is almost normal multidimensional distribution), then random scattering of $\Omega_{\Sigma}\left(\boldsymbol{\beta}\right)$ near to the minimum is close to "chi-square" distribution with a number of degrees of freedom equal to difference between a number of grouped summands in (13) and a number of dimensions in the vector $\boldsymbol{\beta}$.

Bearing in mind that in practice parametric dependence of $L\left(Data \mid \boldsymbol{\beta}\right)$ and $\psi\left(\boldsymbol{\beta} \mid Data\right)$ likelihoods can turn out to be far from multidimensional normal distribution, it seems advisable to complete an estimation algorithm as per a logic following the sequential continuation within Bayesian approach. It means transition to interval estimates based on multidimensional joint distributions (7, 8). However, together with sufficient strictness, Bayesian approach is highly labor-consuming and prone to accumulating computational errors related to direct calculation of multidimensional integrals with participating probability density $\psi\left(\boldsymbol{\beta} \mid Data\right)$ within the space of parameters $\boldsymbol{\beta}$. Given that, in practice it seems more realistic to apply multiple probability simulation (Monte–Carlo method) since it allows us to average functions or parameters that are being considered, together with assessing their marginal statistic properties, as per a great number of point pseudo-observations (~1 million or more) that comply with the distribution $\psi\left(\boldsymbol{\beta} \mid Data\right)$. This algorithm is also labor-consuming, but still, the most realistic one. It is implemented in practice via several ways and one of them, Gibbs' algorithm, is based on well-known theoretical grounds [28]. Herewith estimates of a distribution center as per maximum likelihood method in case density $\psi\left(\boldsymbol{\beta} \mid Data\right)$ is unimodal can be relevant initial approximation for building up a stochastic sequence of pseudo-observations.

Results obtained via assessing risk rates in a heterogeneous cohort with previously known properties. Let us consider an artificially created epidemiologic register that describes a certain "etalon" cohort with events not being random in it and complying with a previously known model. Choice on imitation data instead of actual ones is preferable in this

case since there is no actual register with risk research results obtained for it certainly coinciding with exact and previously known rates. Since only determinate imitation of stochastic behavior by participants may occur in an artificial cohort, we should expect that zero or almost zero deviation (13) would probably be reached numerically. In other words, we are to give a practical answer to a question whether the examined estimation algorithm has asymptotic convergence. This question concerns not only the examined algorithm but also its prototype, AMFIT algorithm within Epicure software package, although it has never been checked before.

Let us consider a radiation-epidemiologic study with the following underlying characteristics as an example. We construct an imitation sampling made up of distinctly limited homogenous strata that differ as per a gamma-irradiation dose (from 0 to 2 Sv), sex, and age. We consider a cumulative radiation dose to be a risk factor and this dose is a result of single acute even irradiation of the lungs that occurred at an age of 19. Our basic reasons for changes in oncologic mortality are a) conditionally linear growth in such cumulative rate as intensity of a risk of death caused by lung cancer with a growth in a dose that is known and limited as per its value; b) a certain decrease in life expectancy for all irradiated members in the cohort; c) sex as a heterogeneity factor that results in both background risk parameters being different for men and women and differences in sensitivity to radiation. Such cause-and-effect relations are well known due to a series of studies on an actual cohort made of people who survived atomic bombing in Hiroshima and Nagasaki [29, 30]. Figure 2 shows a fragment of a diagram (men) with rough estimates of annual risks as per a typical scheme for a three-factor "etalon cohort" (sex, dose, and age). Overall there were 25,000 individual entries in the database (15,000 men and 10,000 women). Overall number of death cases caused by specific cancer providing the cohort totally died out amounted to 1,118 (995 cases among men and 123 cases among women). Overall period of observation over the cohort amounted to 1,031,414 person-years. If we group the participants as per 14 age intervals, 2 sexes, and 5 dose levels, we obtain $14 \cdot 2 \cdot 5 = 140$ homogenous strata. Only 91 out of them turned out to contain non-zero number of specific cancer cases. Non-empty strata corresponded to 754,106 person-years of observation.

It seems obvious that in case there are no random fluctuations in sampling parameters in this "etalon cohort", dose- and age-trends in Figure 2 are to be visible to the naked eye. Simultaneously quantitative regularities for background risk rates fully correspond to population ones [4]. An expert who studies risks should "see" all the preset detailing. So what results can be yielded due to the suggested algorithm and its closest prototype?

The suggested procedure for estimating changes in cumulative and intensive rates turns out to be quite efficient for solving the task; we can see it in Figures 3 and 4, both for women sub-cohort and men sub-cohort within the unified parametric model ($\boldsymbol{\beta} \in R^8$).

Calculated minimum deviation $\Omega_{\min}$ for the computed extreme solution amounted to only 0.40 units in this case given a significant increase in component "observations" as per Kaplan – Meier and a growth in a number of degrees of freedom in comparison with traditional preliminary grouping with applying 5-year age intervals (Figure 2). It, if recalculated

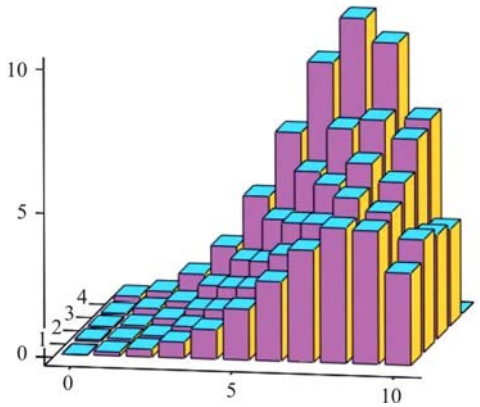

Figure 2. Three-dimensional diagram "age-dose-rate" for men in the "etalon cohort". Horizontal axes show age strata (14) and dose strata (5). Vertical axis shows rough estimates of specific mortality rate within a stratum (‰ per year)

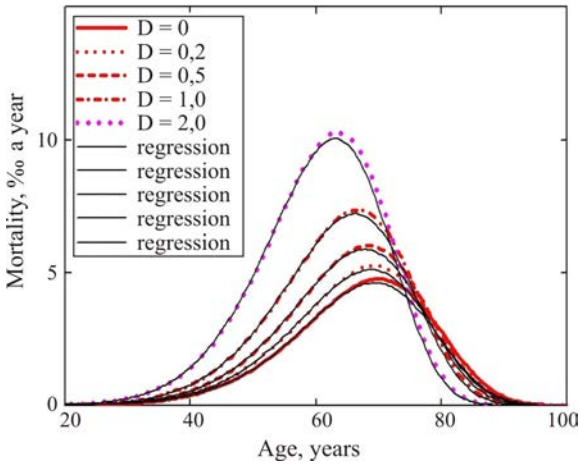

Figure 3. Dynamic dependences for specific mortality rates among men in "etalon cohort" preset and estimated as per binomial regression algorithm

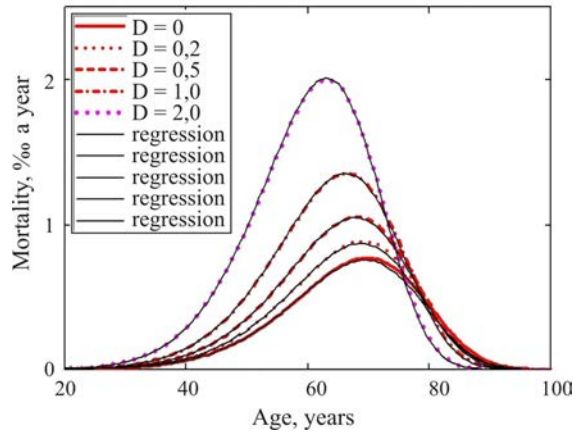

Figure 4. Dynamic dependences for specific mortality rates among women in "etalon cohort" preset and estimated as per binomial regression algorithm

per 1 case out of 1,118, approximately corresponds to visually observable standard mean-root square deviation in estimated "hazard" rates being about $\sqrt{0.40/1118} = 0.019 = 1.9\,\%$. Residual deviation $\Omega_{\min}$ didn't reach exact zero and it indicates that it is impossible to overcome discrete nature of entries in the sampling database in comparison with continuous nature of parameters in an actual general population. However it is hardly possible to further reduce this deviation. Coefficients for a dose trend in cumulative value of the accumulation intensity of the excess lifetime risk and its uncertainty assessed as per Fischer's information matrix

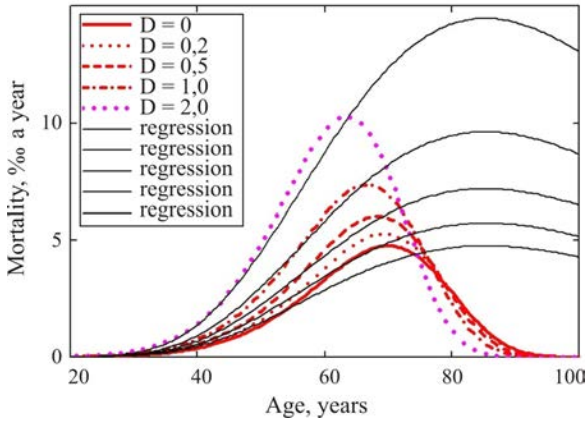

Figure 5. Results obtained via AMFIT-estimates of does and age trends in specific risk compared to its actual behavior for men sub-cohort. Obviously, a typical model [9, 10, 31–33] showing background and examined risks behavior is a source of systemic deviations in all estimates

amounted to $0.49\,Gy^{-1}$ $(95\%\,\mathrm{CI}: 0.24 \ldots 0.99)$ for men and $0.68\,\mathrm{Gy}^{-1}$ $(95\%\,\mathrm{CI}: 0.33 \ldots 1.35)$ for women. We should pay attention to significant uncertainty that still occurs although achieved deviation is extremely low. It occurs due to dimensions of uncertainty area being extremely dependent on function (13) curving close to the extremum and not on its reached minimum value since confidence intervals are estimated in such ways so that countings in the "etalon cohort" were still prone to random fluctuations to the same extent as if they were real ones. It is the basic difference between binomial regression and regression as per least-squares procedures.

Unlike the obtained results, AMFIT procedure, when applied within its typical group of models [9, 10, 31–33], showed significant systemic deviation in annual risks in older ages areas (Figure 5, men sub-cohort). Only ascending parts in the curves can be approximated properly. Also models within this algorithm turned out to fail "to see" probable intersections between group of "hazard" curves related to a certain decrease in life expectancy of irradiated people.

Observed minimum deviation (the formula 13) amounted to 24.3 units for computed extreme solution with a number of degrees of freedom being 85 (a number of strata minus a

number of parameters). Given this number of degrees of freedom, 90 % -interval of expected random minimum deviation occurrence that is seemingly [26] distributed as per "chi-square law" should amount to 64.7 … 107.5. Therefore, the observed value 24.3 is less statistically significant than a typical random value. This observable overfitting practically assuredly indicates that this "etalon cohort" is an artificial one. However, it is rather surprising that $\Omega_{\min}$ deviated from zero significantly; it is obviously due to absurd systemic bias of $h\left(t\middle| D, sex, \boldsymbol{\beta}\right)$ in an area with ages exceeding 75 years and even 100 years for all dose exposures. Actually reached significance $p \sim 10^{-11}$ formally indicates there is extremely low probability that a model will deviate from data, but it corresponds to an actual situation only in an area where risk intensity grows rapidly (Figure 5) but not in a wide range of ages. Since all curves showing annual risks have been biased, the same has happened to dose trends. For example, annual parameter $h\left(t\middle| D, sex, \boldsymbol{\beta}\right)$ for men aged 60 years moved at a rate of changes in a dose being $\approx 82\ \%\ \text{per}\ 1\ \text{Sv}$, and it is almost 3 times higher than a dose trend parameter for cumulative lifetime value determined via the previously examined algorithm under the same conditions. Discrepancy between these two types of relative trends and, consequently, the necessity to distinguish between them has also been mentioned by other authors [34]. Therefore, AMFIT can be prone to overestimating effects produced by irradiation. In some cases use of annual parameters may also result in actual (cumulative) risks being underestimated [31]. Potential basis for such errors is obvious in Figures 3 and 4.

Algorithm for assessing risks in a heterogeneous cohort: discussing advantages and drawbacks. Let us mention the basic aspects in which our algorithm for risk measuring differs from its analogues and primarily from its prototype, AMFIT algorithm. First of all, countings of all examined specific events are considered to be binomial processes in the suggested algorithm, and cumulative and intensive rates are determined on probabilistic basis and not as heuristic values. It allows using risk assessment as a procedure for indirect measurements of continuously-distributed parameter estimates basing on Bayesian approach. As opposed to that, both AMFIT algorithm, and some other algorithms that are not so frequently applied [35–38] are based on point estimate of the whole set of events and smooth approximation of obtained non-smooth empiric distributions within Pearson and Fischer's frequency-discrete statistic paradigm.

These mentioned alternative approaches to risk assessment have obvious but frequently neglected drawbacks together with mathematical simplicity in comparison with process assessment. For example, an actual intensive risk is not either constant or a set of constants as it would be stated within Poisson's statistics. Owing to it, a role played by inaccuracy related to stratification of a heterogeneous cohort as per age is not clear. Too small age intervals can results in cases disappearing in them for cohorts that are small in volume and Poisson's regression functional can lose its extreme properties in that case. Too large intervals, on the contrary, will result in groundless averaging of risk accumulation intensities within an interval. It is difficult to set an optimal width for intervals in advance when performing stratification as per age. A similar drawback of Poisson's regression becomes obvious as a cohort dies out and a number of specific cases tends to zero. What boundary in age distribution should we stop at? Binomial law for event distribution would be more relevant here since Poisson's law arises from it asymptotically as a partial statistical model for rare events.

We should also note that a quality of approximation $h\left(t\middle| D, sex, \boldsymbol{\beta}\right)$ for describing exposure to radiation depends on how well this function operates in case there is no external influence ($D = 0$). Here we speak about this exact reference group which is an integral component in any comparative study. This nuance is often neglected by researchers [32, 33] since they rely on simple models for power-level growth or models with saturation. Formally it means that background life-long cu-

mulative intensity can reach very high values or approach infinity and a specific risk of death due to the examined cause reaches 1. However, there are no such diseases in reality as it can obviously be seen already at a stage when data are being grouped preliminarily (Figure 2). Such data have already been successfully taken into account in models showing intensive risk parameters in dynamics within Bayesian studies on limited samplings with an incomplete period of observation and data losses, for example as per such procedures as "right censored spell models" [39], "cure rate models" [40], "bounded cumulative hazard models" [41]. Moreover, we can justly assume that a shift in estimates in Figure 5 is predominantly related to improper background risk model and not only to the nature of the process being neglected and a statistic law being selected incorrectly. We should also note that such a parameter as *p-value* that is determined via testing likelihoods is not always a reliable reference point when approximation quality is assessed in a multidimensional case.

All the above mentioned doesn't mean that the newly suggested procedure for risk assessment is an ideal researcher's tool. Recall that it is based on such cohort stratification that doesn't provide for cohort members going from one stratum into another during an observation period. Should such transition appear, strata can't be considered independent and it means that functional of probabilistic estimation should be built on other grounds. Another vulnerability is that factor space is not completely covered with available observations; this vulnerability is typical for any empiric sampling. CONSORT standards [42] cover any sampling studies. Even if there is a strong correlation in one pair of factors, effects produced by one of them can be disguised by seeming effects produced by another. Due to that, if we want to assess processes successfully, we should either work with sufficiently large and diverse samplings, or we should examine correlations between risks and factors via a controlled experiment.

It is important to note that this paper doesn't promote any theory in the sphere of analytical epidemiology. We merely suggest one instrument for practical analysis within a "dose – time – risk" mental scheme instead of a conventional limited approach "dose – risk" since the latter imposes interpreting harm caused by external influences only as simple one-dimension dependence in a plain two-axis graph. Indeed, a simplifying "dose – risk" concept can create artifacts in a form of false trends that are hard to explain. Interpretation of a radiation-oncologic trend which we took form the document [43] is a good example.

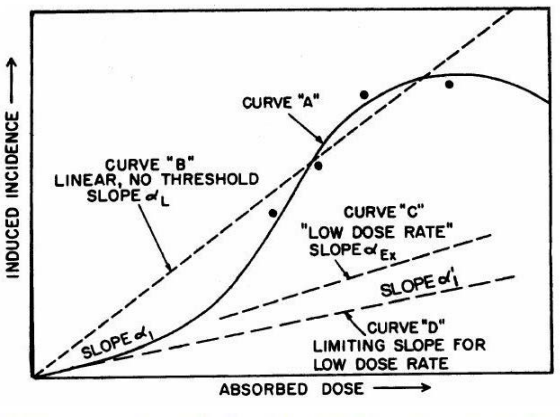

Figure 6. Schematic behavior of excessive risk depending on a radiation exposure dose as expected by BEIR VII experts. Taken from the report [43]

This graph obviously shows a non-linear response in excessive intensity of specific events with a clear non-monotonous drop in an area of large doses. This trend can't be plausibly explained either within a well-known liner non-threshold model or within well-grounded linear models. However, it should be noted that if "intensity" is used instead of "risk", this graph can be linked to only one age group in the cohort. In this case it is easy to reveal that a dose trend shown in Figures 3 and 4 looks exactly like this within age range from 60 to 65 years with a typical drop and even a change in a sign of "excessive" effects in older age groups. And here cumulative risk grows only monotonously with a growth in a cumulative dose. Here we should also remember that a well-known linear-non-threshold "dose – effect" model by N.V. Timofeev-Ressovskiy and K.G. Zimmer [44, 45] was developed exactly

for a pair of cumulative values, an analogue of the Nelson – Aalen estimator being one of them long before this estimator was invented. An author who developed a well-known "effective dose" concept [46] had similar opinions about a sphere where cumulative parameters could be successfully applied.

Conclusion. Therefore, it seems quite promising to analyze dynamics of specific events occurrence in a heterogeneous cohort combined with Bayesian methodology for risk assessment provided that researchers have detailed information about cohort members collected during a sufficiently long period of time or even in a life-long observation and complete and comprehensive description of individual risk factors. Applied computation technique is within conventional epidemiologic procedures for health risk assessment since it combines application of annual group risk rates together with cumulative ones. It has been shown that when experts try to predict damage caused by external influence on people's health within conventional "dose – effect" mental schemes, they should preferably rely on a combination of cumulative doses and cumulative risks or their descriptive analogues (effects).

But at the same time we can't fail to mention that using the described parametric version of Bayesian procedures is rather labor-consuming. This drawback can be partially overcome only via creating relevant software[6] that is able to provide automatic tools for grouping data, selecting models, searching for extreme solutions, and modeling statistic uncertainty of Bayesian estimates.

Funding. The procedure was created within a research program funded by the Federal Medical Biological Agency ("Consequences – 2020").

Conflict of interests. The authors declare no conflict interest concerning this publication.

References

1. Onishchenko G.G., Zaitseva N.V., May I.V. [et al.]. Health risk analysis in the strategy of state social and economical development: monograph. In: G.G. Onishchenko, N.V. Zaitseva eds. Moscow, Perm, Publishing house of the Perm National Research Polytechnic University Publ., 2014, 738 p. (in Russian).

2. Commonwealth of Australia, 2012. Environmental Health Risk Assessment. Guideline for assessing human health risk from environmental hazards: Glossary. Commonwealth of Australia, 2012, 244 p.

3. Publikatsiya 103 Mezhdunarodnoi Komissii po radiatsionnoi zashchite (MKRZ) [Publication No. 103 issued by the International Commission on Radiological Protection (ICRP 103)]. In: M.F. Kiselev, N.K. Shandala eds. Moscow, OOO PKF «Alana» Publ., 2009, 344 p. (in Russian).

4. Zlokachestvennye novoobrazovaniya v Rossii v 2018 godu (zabolevaemost' i smertnost') [Malignant neoplasms in Russia in 2018 (morbidity and mortality)]. In: A.D. Kaprin, V.V. Starinskii, G.V. Petrova eds. Moscow, MNIOI im. P.A. Gertsena Publ., 2019, 250 p. (in Russian).

5. Handbook of epidemiology. In: W. Ahrens, I. Pigeot eds. Switzerland, Springer Publ., 2005, 1617 p. DOI: 10.1007/978-3-540-26577-1

6. Newman J.R. Mathematics of a lady tasting tea by Sir Ronald Fisher. The World of mathematics. Vol. III, Part VIII. New-York, Simon and Schuster Publ., 1956, pp. 1514–1521.

7. Vaupel J.W., Manton K.G., Stallard E. The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography*, 1979, vol. 16, no. 3, pp. 439. DOI: 10.2307/2061224

8. Mikhal'skii A.I., Petrovskii A.M., Yashin A.I. Teoriya otsenivaniya neodnorodnykh populyatsii [Theory of heterogeneous population estimation]. Moscow, Nauka Publ., 1989, 128 p. (in Russian).

9. Preston D.L., Kato H., Kopecky K.J., Fujita S. Technical Report No. 1-86. Life span study report 10. Part 1. Cancer mortality among A-bomb survivors in Hiroshima and Nagasaki, 1950–1982. *RERF*, 1987, no. 111, pp. 151–178.

10. Preston D.L., Cullings H., Suyama A., Funamoto S., Nishi N., Soda M., Mabuchi K., Kodama K. [et al.]. Solid cancer incidence in atomic bomb survivors exposed in utero or as young children. *Journal of the National Cancer Institute*, 2008, vol. 100, no. 6, pp. 428–436. DOI: 10.1093/jnci/djn045

[6] Obesnyuk V.F. A program that manages multi-factor assessment of group health risks as per individual entries in a register containing long-term observations / registered in the patent Office of the RosPatent on December 23, 2020; certificate No. 2020667423.

11. Epicure. The premiere software for risk regression and person-year tabulation. *«EPICURE» Risk Sciences International*. Available at: https://risksciences.com/epicure/ (21.04.2021).

12. Gauss K.F. Izbrannye geodezicheskie sochineniya [Selected geodesic works]. In: G.V. Bagratun, S.G. Sudakov eds. Moscow, IGL Publ., 1957, vol. 1, 153 p. (in Russian).

13. Vaeth M., Pearce D. Calculating excess lifetime risk in relative risk models. *Environmental Health Perspectives*, 1990, vol. 87, pp. 83–94. DOI: 10.1289/ehp.908783

14. Thomas D., Darby S., Fagnani F., Hubert P., Vaeth M., Weiss K. Definition and estimation of lifetime detriment from radiation exposures: principles and methods. *Health Physics*, 1992, vol. 63, no. 3, pp. 259–272. DOI: 10.1097/00004032-199209000-00001

15. Ulanowski A., Kaiser J.C., Schneider U., Walsh L. Lifetime radiation risk of stochastic effects – prospective evaluation for space flight or medicine. *Ann. ICRP*, 2020, vol. 49, no. 1, pp. 200–212. DOI: 10.1177/0146645320956517

16. Ulanowski A., Kaiser J.C., Schneider U., Walsh L. On prognostic estimates of radiation risk in medicine and radiation protection. *Radiat. Environ. Biophys*, 2019, vol. 58, no. 3, pp. 305–319. DOI: 10.1007/s00411-019-00794-1

17. Esteve J., Benhamou E., Raymond L. Statistical methods in cancer research. Descriptive epidemiology. *IARC Scientific Publication*, 1994, vol. IV, no. 128, pp. 313.

18. Sasieni P.D., Shelton J., Ormiston-Smith N., Thomson C.S., Silcocks P.B. What is the lifetime risk of developing cancer? The effect of adjusting for multiple primaries. *Br. J. Cancer*, 2011, vol. 105, no. 3, pp. 460–465. DOI: 10.1038/bjc.2011.250

19. Aalen O., Andersen P.K., Borgan Ø., Gill R.D., Keiding N. History of application of martingales in survival analysis. *Electronic Journal of History of Probability and Statistic*, 2009, vol. 5, no. 1, pp. 1–28.

20. Aalen O., Borgan Ø., Gjessing H. Survival and Event history analysis: A process point of view. New-York, Springer Science + Business Media B.V. Publ., 2008, pp. 539.

21. Grunkemeier G.L., Jin R., Eijkemans M.J.C., Takkenberg J.J.M. Actual and actuarial probabilities of competing risks: apples and lemons. *The Annals of Thoracic Surgery*, 2007, vol. 83, no. 5, pp. 1586–1592. DOI: 10.1016/j.athoracsur.2006.11.044

22. Kaplan E.L., Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* 1958, vol. 53, no. 282, pp. 457–481. DOI: 10.1007/978-1-4612-4380-9_25

23. Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics,* 1972, vol. 14, no. 4, pp. 945–966. DOI: 10.1080/00401706.2000.10485975

24. Bure V.M., Parilina E.M., Rubsha A.I., Svirkina L.V. Survival analysis of medical database of patients with prostate cancer. *Vestnik SPbGU. Seriya 10*, 2014, vol. 10, no. 2, pp. 27–35 (in Russian).

25. Fisher R.A. On the mathematical foundations of theoretical statistics. *Phil. Trans. of the Royal Soc. of London. Series A*, 1922, vol. 222, pp. 309–368. DOI: 10.1098/rsta.1922.0009

26. Wilks S.S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 1938, vol. 9, no. 1, pp. 60–62. DOI: 10.1214/aoms/1177732360

27. Fan J., Hung H., Wong W. Geometric understanding of likelihood ratio statistics. *JASA*, 2000, vol. 95, no. 451, pp. 836–841.

28. Gelfand A.E., Smith A. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 1990, vol. 85, no. 410, pp. 398–409. DOI: 10.1080/01621459.1990.10476213

29. Sources and effects of ionizing radiation. UNSCEAR 1994 report to General Assembly. New-York, United Nations Scientific Committee on the Effects of Atomic Radiation Publ., 1994, 272 p.

30. Effect on ionizing radiation. UNSCEAR 2006. Report to General Assembly. New-York, United Nations Scientific Committee on the Effects of Atomic Radiation Publ., 2008, vol. 1A, 16 p.

31. Finashov L.V., Kuznetsova I.S., Sokol'nikov M.E., Skukovskii S.G. Radiation risk of prostate cancer incidence due to external gamma-exposure in the cohort of «Mayak» PA workers occupationally subjected to prolonged radiation exposure. *Voprosy radiatsionnoi bezopasnosti*, 2020, no. 2, pp. 37–48 (in Russian).

32. Tukov A.R., Shafranskii I.L., Prokhorova O.N., Ziyatdinov M.N. The incidence of cataracts and the radiation risk of their occurrence in liquidators of the Chernobyl accident, workers in the nuclear industry. *Radiatsiya i risk,* 2019, vol. 28, no. 1, pp. 37–46 (in Russian).

33. Kreisheimer M., Sokolnikov M.E., Koshurnikova N.A., Khokhryakov V.F., Romanow S.A., Shilnikova N.S., Okatenko P.V., Nekolla E.A., Kellerer A.M. Lung cancer mortality among nuclear workers of the Mayak facilities in the former Soviet Union. *Radiat. Environ. Biophys*, 2003, vol. 42, no. 2, pp. 129–135. DOI: 10.1007/s00411-003-0198-3

34. Zöllner S., Sokolnikov M.E., Eidemüller M. Beyond two-stage models for lung carcinogenesis in the Mayak workers: implications for plutonium risk. *PLoS ONE*, 2015, vol. 10, no. 5, pp. e0126238. DOI: 10.1371/journal.pone.0126238

35. Demin V.F., Ivanov S.I., Novikov S.M. Common methodology of health risk assessment for impact of different harm sources. *Meditsinskaya radiologiya i radiatsionnaya bezopasnost'*, 2009, vol. 54, no. 1, pp. 5–15 (in Russian).

36. Ivanov V.K, Gorsky A.I., Kashcheev V.V., Maksioutov M.A., Tumanov K.A. Latent period in induction of radiogenic solid tumors in the cohort of emergency workers. *Radiation and Environmental Biophysics*, 2009, vol. 48, no. 3. DOI: 10.1007/s00411-009-0223-2

37. Ivanov V.K., Tsyb A.F., Panfilov A.P., Agapov A.M. Optimizatsiya radiatsionnoi zashchity: «dozovaya matritsa» [How to optimize radiological protection: «dose matrix»]. Moscow, Meditsina Publ., 2006, 304 p. (in Russian).

38. Jacob P., Meckbach R., Sokolnikov M., Khokhryakov V.V., Vasilenko E. Lung cancer risk of Mayak workers: modeling of carcinogenesis and bystander effect. *Radiat. Environ. Biophys*, 2007, vol. 46, no. 4, pp. 383–394. DOI: 10.1007/s00411-007-0117-0

39. Chen M., Ibrahim J., Sinha D. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 1999, vol. 94, no. 447, pp. 909–919. DOI: 10.1080/01621459.1999.10474196

40. Rodrigues J., Balakrishnan N., Cordeiro G., de Castro M. A unified view on lifetime distributions arising from selection mechanisms. *Computational Statistics and Data Analysis*, 2011, vol. 55, no. 12, pp. 3311–3319. DOI: 10.1016/j.csda.2011.06.018

41. Tsodikov A.D., Ibrahim J.G., Yakovlev A.Y. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 2003, vol. 98, no. 464, pp. 1063–1067. DOI: 10.1198/016214503000000001007

42. Moher D., Hopewell S., Schulz K.F., Montori V., Gøtzsche P.C., Devereaux P.J., Elbourne D., Douglas M.E., Altman G. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomized trials. *International Journal of Surgery*, 2012, vol. 10, no. 1, pp. 28–55. DOI: 10.1016/j.ijsu.2011.10.001

43. Health risks from exposure to low levels of ionizing radiation. BEIR VII, phase 2. Washington D.C., Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation Publ., 2006, 406 p.

44. Timofeev-Resovskii N.V., Tsimmer K.G. Teoriya misheni radiobiologicheskogo deistviya (v izlozhenii) [Theory of radiological-biological effects targeting (a presentation)]. *Biosfera*, 2010, vol. 2, no. 3, pp. 432–450 (in Russian).

45. Zimmer K.G. Ergebnisse und Grenzen der treffertheoretischem Deutung von strahlenbiologischen Dosis-Effekt kurven. *Biol. Zent*, 1941, no. 63, pp. 72–107.

46. Jacobi W. The concept of the effective dose – a proposal of the combination of the organ doses. *Radiat. And Environm. Biophys*, 1975, vol. 12, no. 2, pp. 101–109. DOI: 10.1007/BF01328971

Obesnyuk V.F. Group health risk parameters in a heterogeneous cohort. Indirect assessment as per events taken in dynamics. Health Risk Analysis, 2021, no. 2, pp. 18–33. DOI: 10.21668/health.risk/2021.2.02.eng

Received: 26.05.2021
Accepted: 04.06.2021
Published: 30.09.2021