Original research article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Quantification of the influence of risk factors with application to cardiovascular diseases in subjects with type 1 diabetes

Statistical Methods in Medical Research
1–19

DOI: 10.1177/09622802251327680
journals.sagepub.com/home/smm
Sage

Ornella Moro[1], Inger Torhild Gram[2,3], Maja-Lisa Løchen[4], Marit B Veierød[5], Ana Maria Wägner[6] and Giovanni Sebastiani[1,7,8]

Abstract

Future occurrence of a disease can be highly influenced by some specific risk factors. This work presents a comprehensive approach to quantify the event probability as a function of each separate risk factor by means of a parametric model. The proposed methodology is mainly described and applied here in the case of a linear model, but the non-linear case is also addressed. To improve estimation accuracy, three distinct methods are developed and their results are integrated. One of them is Bayesian, based on a non-informative prior. Each of the other two, uses aggregation of sample elements based on their factor values, which is optimized by means of a different specific criterion. For one of these two, optimization is performed by Simulated Annealing. The methodology presented is applicable across various diseases but here we quantify the risk for cardiovascular diseases in subjects with type 1 diabetes. The results obtained combining the three different methods show accurate estimates of cardiovascular risk variation rates for the factors considered. Furthermore, the detection of a biological activation phenomenon for one of the factors is also illustrated. To quantify the performances of the proposed methodology and to compare them with those from a known method used for this type of models, a large simulation study is done, whose results are illustrated here.

Keywords

Risk quantification, risk factor analysis, simulated annealing, doseResponse curve, bayesian statistics

1 Introduction

A relevant problem in medicine is the quantification of the risk to develop a certain disease. In general, the risk of such a pathological condition is influenced by some variables also known as risk factors. For example, for cardiovascular diseases (CVDs), the main risk factors include age, sex, total cholesterol, triglyceride levels, systolic blood pressure, family history of CVD and smoking.[1,2] A risk factor, can be either binary or not. Let us focus on the case of a non-binary risk factor, described here theoretically as a variable whose generic value f is a positive real number in a certain finite interval. The value of f generally varies among the individuals of the population studied and it is assumed to be described by a probability density function $s(\cdot)$. A common way to study the influence of this factor on the risk is through a dose–response curve $g(\cdot)$,

[1]Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche (CNR), Rome, Italy
[2]Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
[3]Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[4]Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[5]Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
[6]Endocrinology and Nutrition Department, Complejo Hospitalario Universitario Insular Materno-Infantil, Instituto de Investigaciones Biomédicas y Sanitarias (IUIBS), Universidad de Las Palmas de Gran Canaria (ULPGC), Las Palmas de Gran Canaria, Spain
[7]Dipartimento di Matematica Guido Castelnuovo, Sapienza Università di Roma, Rome, Italy
[8]Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway

Corresponding author:
Ornella Moro, Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche, Rome, Italy.
Email: o.moro@iac.cnr.it

which represents the probability that the event E considered happens conditional to having the value f:

$$g(f) := \mathcal{P}(E = 1 \mid f) \tag{1}$$

Different types of models $g(\cdot)$ can be adopted.[3] One of the most frequently used is an s-type function, e.g. logistic, which includes both an initial activation and a final saturation phenomena, and an approximately linear part in the middle:

$$g(f) = a \cdot f + b \tag{2}$$

In the above model, the proportionality constant a, the most important of the two parameters, is the rate of increase (decrease for protective factors) of the probability risk corresponding to an increase of the risk factor by a unit. In this work, we mostly adopt the model in equation (2). However, as shown later, the proposed methodology can also be applied with minor modifications in case a non-linear model is adopted, as shown for the case of

$$g(f) = \exp(a \cdot f + b) \tag{3}$$

which is applied here for one of the considered factors.

A statistical approach known as 'Probit method' was developed by Bliss in 1934[4] for the situation considered here, i.e. a binary dependent variable modeled by a function (linear for example) of an independent one. An extension of this work was made by Fisher one year later,[5] which included an iterative method for parameter estimation by maximizing the likelihood (ML) of the data. In this work, we use the approach of Fisher as a term of comparison for the simulation results obtained by the proposed methodology. The ML method is simple and has good theoretical properties, one of the most important being the estimator's asymptotic normality, which allows to compute confidence intervals.[6] However, the solution provided may be inaccurate in case the maximum is reached on a plateau. The methodology proposed here consists of a combination of three different methods. Unlike the known method, for which many public software are already available, the three methods require a certain coding effort. Furthermore, they are also more computationally intensive. However the combination of three different methods with different features is expected to provide a more reliable and consistent estimation of the parameters. Finally, one of our three methods, the Bayesian one, includes in the model the likelihood function used in the approach of Fisher, but also some additional 'a priori' information.

The first method considered here is developed within the Bayesian framework.[7] As mentioned above, together with the likelihood function, it uses a *non-informative prior* on the parameters, which are considered random. We could then proceed similarly to ML, by looking at the parameter value which maximizes its "a posteriori" probability given the measured data. However, to avoid the same problem of ML, instead of considering the mode, we focus on the mean. Following Bayesian approach, confidence intervals can also be computed. The other two methods are based on aggregation of individuals depending on their values of f, with individuals of the same group having close values of f. The risk is then estimated in each group on the basis of the events that happened to individuals belonging to it. In one of these two methods, the division of the range of values of f in sub-intervals (inducing groups of individuals), is based on a criterion imposing equality of the variances of risk estimators in different sub-intervals. After the optimal division is found, the parameters a and b are estimated by a least squares procedure. In the other method, the optimal division is based on a criterion minimizing the sum of squared errors between the theoretical and empirical risk estimates in different sub-intervals. Differently from the previous method, the optimization is now not straightforward and we perform it by using the Simulated Annealing algorithm.[8,9] In addition, we propose a method for the detection of a region to be excluded from the analysis. This happens when data do not provide evidence against the hypothesis that the risk does not vary in the most left or most right part of the empirical range of the factor. The method incorporates two additional conditions. If also both of them are fulfilled, we claim that this is consistent with the presence of a real biological activation phenomenon.

As an alternative, instead of analyzing risk factors separately, multiple factors linear regression can be performed, as done in validated risk calculators, mostly based on the Cox Proportional Hazard (CPH) model.[10] Existing risk calculators, such as the NORRISK2,[11] which are based on CPH model, have been analyzed and results show good risk prediction performance both in terms of sensitivity and specificity (AUC $\simeq$ 0.8). In this way, we can "adjust" the estimation of the rate of a factor, accounting for the influences of the other covariates on the risk. This same approach is applied also for other two widely used risk calculators for CVD events, which are the SCORE,[12] for the general population, and the STENO1,[13] for people with type 1 diabetes (T1D). However, in this way we do not take into account possible dependencies between factors, which happen in practice. This may lead to a wrong conclusion regarding the relative influence of different factors. We illustrate it by the example included in A. Some consequences of ignoring risk factor dependencies are shown by the examples in Figure 1, relative to data from a longitudinal study of CVD events in individuals with T1D,[14–16] which is intensively used here.

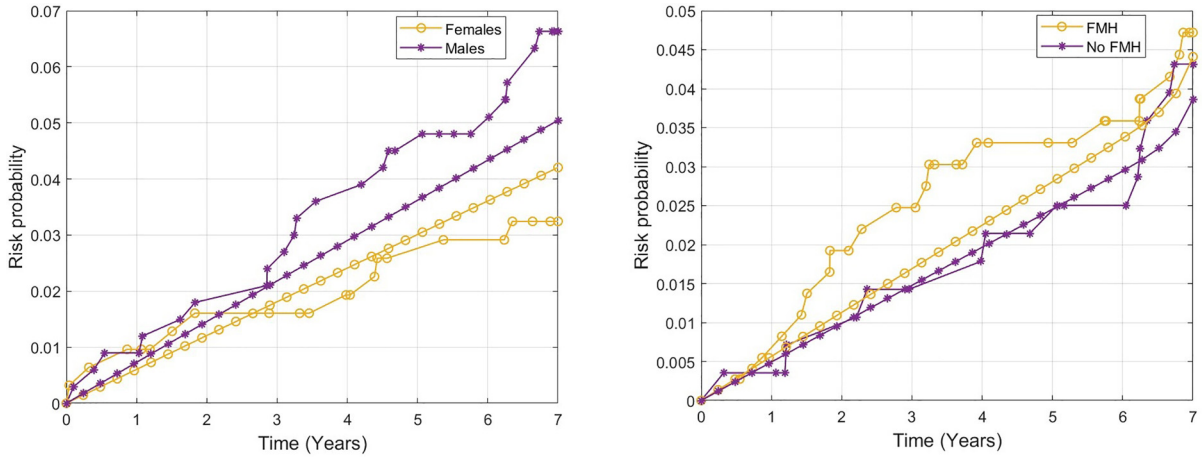

Figure 1. In the two panels, the empirical risk curves obtained by the Kaplan Meier estimator[17] from the two sub-samples of males and females (left panel) and of individuals with and without family history of hypertension (right panel) are shown. The superimposed theoretical curves (smooth lines) are estimated by means of the Cox Proportional Hazard model.

The figure shows the empirical risk probability curves obtained by the Kaplan Meier (KM) estimator[17] for males and females and for individuals with and without family history of hypertension (parents, grandparents, siblings or children) of the longitudinal Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) study for CVD events in individuals with T1D. The theoretical curves are estimated by setting the two values of the binary factors sex and family history after parameter estimation of a model including many risk factors, but they do not fit the empirical ones well. We finally notice that when association between two factors is demonstrated, it does not imply causal relationship. However, in some cases, it is possible to establish a causal relationship in a straightforward way, like for smoking that may influence systolic blood pressure.[18–20] In other situations, this is also possible based on knowledge from medical experts, e.g. the levels of high and low density lipoprotein (LDL) cholesterol may depend on triglyceride level.[21]

Performing dose–response analysis for a risk factor on its own has one relevant advantage. In fact, the rate of increase of a factor takes into account the indirect effects on the event E of other factors that are dependent causally on it. This is useful in practice as the reduction of the value of a modifiable factor may induce changes in the values of other factors, with consequences on the risk. Following this approach, we do not eliminate the effects of factors correlated with E, but not dependent causally on the one considered. However, we provide solutions to cope with this issue, as described later in the discussion.

Here, the methodology is mainly tailored for a linear dose–response model. Of course, in practice, a more complex model may be needed to describe the relationship between risk and factors. The appropriateness of any given model, e.g. the linear one considered here, should be assessed either theoretically or empirically. This is true for any study of this type, included those cited above, based on the CPH model which is also a linear one. However, in this particular study, our focus is mainly on a innovative methodology to estimate the parameters of a given assumed model. Nevetheless, the method can cope also with the case of non-linear relationships. This is done here for one of the factors, for which an exponential model is assumed.

The proposed methodology is general. However, here we illustrate the results of its application for the quantification of the risk of CVD events in individuals with T1D. The results described include accurate estimates of the risk variation rates for some of the main non-binary factors. In addition, the results obtained suggest the presence of a real activation phenomenon for one of the considered risk factors. Finally, to quantify the performances of the proposed methodology, we also include a large study on realistic simulated data.

2 The proposed methodology

We describe here the proposed methodology to provide accurate estimation of dose–response curves. Although this approach can be adopted also for a non-linear function $g(\cdot)$, we now focus on the linear case. However, later in this section we will briefly describe the modifications needed when adopting a non-linear function $g(\cdot)$. As anticipated before, we combine the solutions from three different methods by averaging them. Possibly, the range of analysis $[f_L, f_R]$ is equal to the

empirical one in the whole data sample analyzed. However, in some cases, the range of analysis can be strictly contained in the empirical one, as it will be explained after the description of the three methods. For convenience, we assume that the value of the independent variable f of the linear function in equation (2) is the measured value of the factor to which is subtracted f_L.

2.1 The Bayesian method

Here, we follow the Bayesian approach.[7] We do not describe the linear function $g(\cdot)$ by the two above parameters a and b. Instead, we consider the two parameters g_L and g_R equal to the values of $g(\cdot)$ in the left and right points of the range of the values of the factor where the analysis is performed, that is $g_L = g(f_L)$ and $g_R = g(f_R)$. For a risk (protective) factor we have that $g_L < g_R$ ($g_L > g_R$). We notice that g_L and g_R, which we want to estimate, must be in the interval $[0, 1]$. After the estimation of g_L and g_R, the value of a is obtained by the formula $a = (g_R - g_L)/(f_R - f_L)$.

2.1.1 The statistical model

Let us now consider the likelihood $\mathcal{L}(Y)$ of a sample of n i.i.d. measurements $Y := Y_1, \ldots, Y_n$, given a generic value of the pair (g_L, g_R), as follows

$$\mathcal{L}(Y) := \mathcal{P}(Y|g) = \prod_{i=1}^{n} \mathcal{P}(Y_i|g_L, g_R) = \prod_{i=1}^{n} g(f_i)^{Y_i} \cdot (1 - g(f_i))^{1-Y_i}$$

where f_i is the value of the factor for the ith individual and $Y_i = 1$ when the event E happened, and zero otherwise. In the above formula the dependence of the linear model $g(\cdot)$ on the parameters g_L and g_R has been dropped for simplicity of notation.

Following the Bayesian approach, we consider a probabilistic "a priori" model $\mathcal{P}(G) = \mathcal{P}(G_L, G_R)$ for the parameters, which are considered random variables and indicated by (G_L, G_R). This model is the product of two uniform distributions on G_L and G_R separately. The support of the distribution of G_L (G_R) is $[0, \bar{g}]$ ($[\bar{g}, 1]$), where $\bar{g}$ is the probability of the event E obtained by applying the law of total probability:

$$\bar{g} = \frac{\int_{f_L}^{f_R} g(f) \cdot s(f)\, df}{\int_{f_L}^{f_R} s(f)\, df} \tag{4}$$

where $s(f)$ is the probability density function of the considered factor (whose integral in the factor interval is 1). In fact, both G_L and G_R must be between zero and one. In addition, in case of a genuine risk factor, we have that $g(\cdot)$ is a non decreasing function, therefore it is easy to see how $g_L \leq \bar{g}$ and $\bar{g} \leq g_R$. We can proceed similarly for a protective factor, for which we obtain that $\bar{g} \leq g_L$ and $g_R \leq \bar{g}$. The quantity $\bar{g}$ is estimated here using the sample mean as $\sum_{i=1}^{n} Y_i/n$. We notice that this approach can also be used for a non-linear function $g(\cdot)$, as it will be illustrated later. By means of the Bayes theorem, "a posteriori" distribution is derived

$$\mathcal{P}(G|Y) \propto \mathcal{P}(Y|G)\, \mathcal{P}(G)$$

2.1.2 Parameter estimation

Here, to estimate $a = a(G) = (G_L - G_R)/(f_L - f_R)$, we use its expected value according to the "a posteriori" distribution for G. Very often, in Bayesian statistics this expected value is not known in explicit form as a function of the data (and eventually of a few "hyper-parameters" of the prior distribution). A possible approach is to approximate it with high accuracy by stochastic simulation,[22] for example using the Metropolis algorithm.[23] We notice that, in the above expression of the posterior distribution, the proportionality constant is missing. In fact, when using Metropolis algorithm, we do not need it, as they only use ratios of the distribution. This is a very nice feature, since the exact calculation of the normalizing constant is very often impossible. To assess that the Metropolis algorithm has reached convergence, a statistical test can be applied.[24] However, since only the two variables G_L and G_R are involved here, the numerical computation of the two involved integrals can be done, allowing us to directly estimate the posterior mean. To do that, the expected value of a, expressed as a function of G as $a = (G_L - G_R)/(f_L - f_R)$, is approximated here numerically using a finite bidimensional grid on the support $[0, \bar{g}] \times [\bar{g}, 1]$. The posterior weights are computed on the same grid. In a similar way, we compute

numerically the Bayesian variance of a. The whole approach results particularly convenient here to reduce computational complexity. This makes it possible to perform a large simulation study for quantifying the performances of this method.

2.2 The methods based on data aggregation

In this case, the different subjects of the sample are aggregated in groups, with individuals of the same group having "close" values of the considered factor. For a given value k of the number of groups, we partition the range of analysis $[f_L, f_R]$ in k non intersecting sub-intervals $I_1, \ldots, I_k$ possibly with different lengths. The two methods aggregate data following different criteria, as described below.

Given any sub-interval I_i, let us consider the subset of individuals whose factor values belong to it and let us denote by n_i its size. The number $\tilde{n}_i$ of individuals, among the n_i considered, who have experienced the event E, follows a binomial distribution with parameters (n_i, p_i) where

$$p_i = \frac{\int_{I_i} g(f) \cdot s(f)\, df}{\int_{I_i} s(f)\, df} \tag{5}$$

and $s(f)$ is, as above, the probability density function of the considered factor. Based on the linear model assumed for the function $g(\cdot)$, it is easy to show that

$$p_i = a \frac{\int_{I_i} f \cdot s(f)\, df}{\int_{I_i} s(f)\, df} + b \tag{6}$$

where the ratio of integrals above represents the expected value of f conditioned on being in the sub-interval I_i. From equation (6), given a partition of the range, we can compute the value p_i by $a \cdot \bar{f}_i + b$, where $\bar{f}_i$ is approximated by the arithmetic average of the factor values in the individuals belonging to the interval I_i. The values of a and b are then estimated by minimizing the sum of squares of the differences $\hat{p}_i - (a \cdot \bar{f}_i + b)$, where $\hat{p}_i = \tilde{n}_i / n_i$.

In the followings, we describe the two non-Bayesian methods proposed here including the criteria to find the optimal partition of the range of analysis. To take into account the influence of the number k of partition sub-intervals on the estimates, we repeat the whole procedure for different values of k and we average the estimates obtained. We notice that we could proceed similarly for the case of a non-linear function $g(\cdot)$. In fact, we could approximate it by a suitable polynomial with coefficient vector $\boldsymbol{\theta}$. Then, the r.h.s. of equation (5) would become a linear combination of the conditional moments of the variable f, up to the order of the chosen polynomial.

2.2.1 The method based on estimator variance

The expected value of $\tilde{n}_i$ is $n_i \cdot p_i$. Therefore, the estimator $\hat{p}_i$ is an unbiased estimator for p_i, whose variance is given by

$$v_i = \frac{p_i(1 - p_i)}{n_i} \tag{7}$$

For small values of p_i, the involved variance reduces to p_i / n_i, and the ratio between it and p_i becomes $1/n_i$. To find the division of the range of analysis in sub-intervals, we adopt the criterion of equality of the ratios between the estimator variances in different sub-intervals and the corresponding estimator expected values. This implies that all n_i are equal to each other. Therefore, each of the k frequencies n_i / n is equal to $1/k$. We now show how we find the right points of the sub-intervals. From the values of f in the n individuals of the sample, we compute the empirical cumulative distribution $\hat{F}$ of that factor. By using the Nadaraya-Watson kernel regression[25,26] we can interpolate it in any point of the range by a function $F(\cdot)$. Then, we find the right point ξ_i of the ith sub-interval by imposing that $F(\xi_i) = i/k$, $i = 1, \ldots, k$. Once the division of the range in k sub-intervals is obtained, the quantities a and b are computed as described above by minimizing the sum of squares of the differences $\hat{p}_i - (a \cdot \bar{f}_i + b)$.

2.2.2 The method based on unexplained fluctuations

Here we follow a different criterion for data aggregation. In fact, the minimization of the above sum of squares differences (unexplained fluctuations) is used also to find the optimal partition. As candidates of the range division in sub-intervals, we only consider those cases where in each sub-interval there are at least 100 individuals with at least one of them experienced

the event E. The minimization is performed here numerically by means of the Simulated Annealing algorithm with a geometrically decreasing temperature schedule with base close to one.[8,9] The sub-intervals $I_1, \ldots, I_k$ can be identified by the ordered sequence of their $k-1$ points belonging to the interior of the range. At each iteration, we select consecutively each of those point and we propose a new location for it. The location is chosen uniformly at random between the most-left and the most-right points, ensuring that in each subintervals there are at least 100 observations.

2.3 The non-linear case

We now give some details for the application of the proposed methodology in the case of a non-linear model for the function $g(\cdot)$. As an example, we consider the case of an exponential model, as in equation (3), which is applied here for HBA1C. For the Bayesian method, we proceed similarly to the case of a linear function $g(\cdot)$ by performing model re-parametrization. However, this time, the expression of the new parameters (g_L, g_R) as a function of a and b is $G_L = e^b$ and $G_R = \exp\{a(f_R - f_L) + b\}$. By inverting the above mapping, we have $a = \frac{1}{G_R - G_L} \ln\left(\frac{G_R}{G_L}\right)$ and $b = \ln(G_L)$. These expressions are inserted in the likelihood function used for the Bayesian method. Finally, calculating the above inverse function on the estimated values of (g_L, g_R), provides estimates for a and b. When applying the other two methods, instead of using the procedure as in the end of Section 2.2, we proceed as for the linear case, after replacing the quantities $\hat{p}_i$ by their logarithm $\ln(\hat{p}_i)$.

2.4 Range of analysis

As anticipated in the beginning, in some cases, the range of analysis $[f_L, f_R]$ is strictly contained in the whole empirical one. This happens when, below a certain threshold f_A, data do not provide evidence that the risk probability increases as the factor value does. In this case, the left point f_L is replaced by f_A. For a protective factor, it may happen that the reduction of the risk, increasing the factor values, does not happen above f_A. Then, the right point f_R is replaced by f_A.

Now, details of the developed method to detect this phenomenon and eventually to find the value f_A are provided. We focus on a risk factor, like systolic blood pressure (SBP). We describe theoretically this situation by assuming $\mathcal{P}(E = 1 \mid f)$ to be equal to b for measured value of f smaller or equal than f_A, while afterwards it follows the model in equation (2), where f is replaced by $f - f_A$. For a protective factor, like high density lipoprotein (HDL), this may happen for factor values larger than f_A, where again the risk is assumed to be constant. As done for the estimation of the rate a, also here, our method combines two different procedures described below. It could also be easily adapted to deal with cases where the above phenomenon happens, for a standard factor, above a certain threshold. However, here we apply it only in the way described above. This is because the application of our two non Bayesian methods on the whole factors ranges show, for standard factors, visual evidence of flatness only on the most left part of the ranges. Conversely, for the protective factor HDL, this only happens on the most right part of its range. When the range of analysis is reduced, the three methods to estimate the rate a are slightly modified, as described later.

2.4.1 First procedure

Here, we look at two adjacent sub-intervals of equal length $I_L := [f_0 - \Delta, f_0]$ and $I_R := [f_0, f_0 + \Delta]$ contained in the empirical range. The value of Δ must be selected from data. As it will be explained later, we will use different values for Δ. For the moment, we consider it fixed. We consider the two sub-samples of individuals whose factor values belong to each sub-interval and we compute the estimated values p_L and p_R by means of the corresponding KM curves at last time. We then let the two windows move together by increasing the value of f_0 from the minimum allowed one. If both windows are on the left of f_A, the value of the function $g(\cdot)$ is equal to b in all the points of I_L and I_R. Therefore, by equation (6) it is easy to see that $p_L = p_R = b$, and $p_R - p_L = 0$. When f_A coincides with the common point f_0 of the two sub-intervals, p_L still remains equal to b, but by equation (6) it is easy to see $p_R = a \cdot \bar{f}_R + b$. Then we have $p_R - p_L = a \cdot \bar{f}_R > 0$. Analogously, in case both sub-intervals are on the right of f_A, it follows $p_R - p_L = a \cdot (\bar{f}_R - \bar{f}_L) > 0$. This suggests to set f_A equal to the first value of f_0 (if it exists) from which the value of the KM curve at last point corresponding to I_R is above the other one. To increase the statistical significance of the procedure, we require the condition being fulfilled in the last half time interval of the KM curves.

2.4.2 Second procedure

In this case, we only consider a small sub-interval $[f_0, f_0 + \Delta]$ contained in the empirical range, and the sub-sample of individuals whose factor values belong to it. This sub-sample is further divided in two depending if its generic individual experienced or not the event E. If there are events in the sub-interval, let us denote by $\boldsymbol{f}_E$ and $\boldsymbol{f}_{\bar{E}}$ the corresponding sets of values of that factor in the two sub-samples. We now consider any sub-interval lying on the left of f_A ($f_0 + \Delta < f_A$). Then,

the distribution of f, as well as its expected value, is identical for both cases E and $\overline{E}$. Therefore, the difference $\mu_E - \mu_{\overline{E}}$ of their expected values is zero. When the sub-interval contains f_A ($f_0 \leq f_A \leq f_0 + \Delta$), the linear model is active in a part of the sub-interval and $\mu_E - \mu_{\overline{E}}$ is expected to be larger than zero. In the Appendix B, we prove that $\mu_E - \mu_{\overline{E}}$ is larger than zero in case the sub-interval is on the right of f_A ($f_A < f_0$). Of course $\mu_E - \mu_{\overline{E}}$ is equal to the expected value of the difference between the sample means $d(\boldsymbol{f}_E, \boldsymbol{f}_{\overline{E}}) = \langle \boldsymbol{f}_E \rangle - \langle \boldsymbol{f}_{\overline{E}} \rangle$. Therefore, we expect that, if the sub-interval moves from left to right but remaining on the left of f_A, the value of $d(\boldsymbol{f}_E, \boldsymbol{f}_{\overline{E}})$ fluctuates around zero. Instead, when the sub-interval is on the right of f_A, the fluctuations are around a value larger than zero. Hence, we assign to f_A the value of the left point f_0 such that the maximum of $d(\boldsymbol{f}_E, \boldsymbol{f}_{\overline{E}})$ is reached at first, with its value being significantly larger than zero. To test the hypothesis that this maximum difference is statistically larger than zero, we compute the Wilcoxon rank sum statistics[27] $W(\boldsymbol{f}_E, \boldsymbol{f}_{\overline{E}})$ and we use a significance level equal to 0.05. This statistic is commonly used to test (non-parametrically) the null hypothesis of equality between two expected values. The distribution of the test statistic W under the null hypothesis is in general computed either exactly for samples of small size, or asymptotically.[6]

2.4.3 Combination of the two procedures

In practice, for both procedures, we consider a finite number of positions of f_0 from the smallest value f_L with increments of a certain fixed amount δ. We perform the analysis with different values of Δ. Analogously, for a protective factor (here HDL), we apply the same procedure but moving from right to left. The value of δ is estimated from data. Following a conservative principle, we say that there is a region to be excluded from the analysis if, at least one of the two procedures detect it for one or more values of Δ. Based on the same principle, we set f_A equal to the maximum of all values eventually found for it. Of course, for a protective one the maximum is replaced by the minimum. To have confirmation of the goodness of the estimates of f_A, we perform the following test. We divide the interval (f_L, f_A) ((f_A, f_R) for HDL) in two parts with equal number of individuals with factor values in one of them or the other. Then, we test the hypothesis that the risk probabilities for individuals corresponding to the two sub-intervals are equal. For doing that, we use a χ^2 test modified for small numbers.[28]

The method proposed suggests sometimes the reliable application of the linear (or non-linear) model only for values of the factor larger than f_A. However, in some cases, we can go further than that and conclude that a real biological phenomenon exists. There is evidence of this when the estimated value of f_A for both procedures is close to the mode of the probability density function $s(\cdot)$ of the factor, and, for the first procedure, the empirical value of $p_R - p_L$ fluctuates around zero for a large set of sub-interval configurations lying on the left of f_A (compared to the standard deviation of the factor). In fact, when f_A is close to the mode of $s(\cdot)$, we are in the best condition to observe a positive difference of the empirical value of $p_R - p_L$. This can be understood by looking at equation (7) for the variances of p_L and p_R. Indeed, when this condition is fulfilled, the value of the denominator n_i, that is the expected value of the number of individuals in I_R and I_L, is close to the maximum. As it will be seen later, here we can conclude that a real activation phenomenon exists for HDL.

2.4.4 Method modifications

We now describe some small modifications of the three methods in Sections 2.1–2.2 in case when a region has been detected by the combination of the above procedures. For the first method, the information from individuals with factor values smaller (larger for a protective factor) than f_A is not excluded. Indeed, the values f_i for all elements of the data sample with factor values smaller (larger for a protective factor) than f_A are set equal to f_A. The analysis is then performed as described above on the data sample so modified.

For the other two methods, we do not consider all the individuals with factor values below (above for a protective factor) f_A, but only a subset of them with values belonging to a sub-interval I_0 with length Δ and right (left) point f_A. In particular, for the second method, we apply the described procedure to find the division of the range of analysis $[f_A, f_R]$ by considering only the individuals whose values of f belong to that range. After the division has been performed and the pairs $(\bar{f}_i, \hat{p}_i)$ are computed, we add to them the pair $(f_A, \hat{p}_0)$, where $\hat{p}_0$ is the risk estimated in the subset of individuals whose factor values belong to I_0. We proceed in a slightly different way for the third method. In this case, the pair $(f_A, \hat{p}_0)$ is added from the beginning and therefore also influences the optimal division of the range $[f_A, f_R]$. The computation of the final estimate of the slope a and its standard deviation for both methods is performed as follows. We average the estimates of a for different values of Δ in combination with different values of k. From the simulation study, we find a high correlation between those different estimated values of a. Therefore, we compute the standard deviation of the final average as an upper bound by considering each correlation equal to one. The resulting value of the above standard deviation corresponds to the arithmetic average of those from the different combinations of Δ and k. For a given choice of Δ and k, the value of a, obtained by a least squares procedure, is a linear combination of the relative frequencies in each sub-interval with weights depending on the corresponding average of the factor values. This is used to compute the standard deviation of a for each combination.

Table 1. The risk factors included in this analysis. The units used for the different factors, the corresponding median values, modes and measured ranges are reported.

Acronyms	Risk factor (units)	Median value	Mode	Measured range
SBP	Systolic blood pressure (mmHg)	119	115	[89, 172]
LDL	LDL cholesterol (mg/dL)	113	107	[28, 255]
HDL	HDL cholesterol (mg/dL)	55	50	[28, 112]
TRG	Ttriglyceride concentration (mg/dL)	88	54	[20, 635]
HBA1C	HBA1C level (%)	8	7.7	[5, 13]

3 Application

In the following section, we first report the results obtained applying the proposed methodology to a large real dataset from people with T1D. Then, we describe results of the application to a realistic simulated dataset, in order to quantify the performances. The known method used for comparison is applied to the same dataset.

3.1 Application to real data

3.1.1 Real data

The following analysis refers to the information available from a United States National Institutes of Health database of CVD events for individuals with T1D. The study consists of two phases: The DCCT which ends at year 10, when the EDIC observational study starts. The database contains data of 20 years[14–16] of the EDIC study since its start. For each individual of the sample, either the time of the event (if it occurred) or the censoring time (end of the study or withdrawal from it, death) was recorded.[15,16]

A large number of risk factors for CVD are recorded each year of the study, including the non-binary variables considered here and shown in Table 1. We notice that all the non-binary risk factors we include are not directly modifiable as instead is, for example, smoking. However, they are indirectly modifiable through lifestyle changes (e.g. the introduction of specific dietary restrictions or practicing physical activity) and/or the use of medication. Here, we consider as "event" E the first outcome of a CVD episode of any possible type included in the EDIC study.[14–16]

Among all the continuous risk factors available in the considered database, the subset used here is chosen with reference to the two widely used risk calculators NORRISK2 by Selmer et al.[11] and STENO1 by Vistisen et al.[13] The SCORE2, also well known and used, presents only categorical variables and therefore is not considered here. The NORRISK2 is a validated Norwegian tool for the prediction of 10-year acute risk of incident MI or cerebral stroke in individual aged 40–75 years. The STENO1 is an online tool to predict 5 and 10 year risk of CVD for adult people with T1D. In our analysis of the EDIC database, we consider the continuous variables included in the NORRISK2 risk calculator, except for the total cholesterol. We prefer to use LDL cholesterol, which has also been reported to be highly correlated with CVD.[29] Indeed, total cholesterol also includes the antagonist contribution of the protective factor HDL cholesterol. This choice is also motivated by the fact that the STENO1 risk calculator for people with T1D uses this factor. In addition, we also include the triglyceride concentration (TRG).[21] We acknowledge that NORRISK2 is designed for the general population, while the EDIC database used here includes only individuals with T1D. Therefore, we also include in the analysis the level of glycated hemoglobin which is a well-known CVD risk factor for those individuals[30–32] and it is also included in the STENO1.[13] The five risk factors considered here are shown in Table 1.

For each of the considered factors, we average the corresponding values during the first five years of the EDIC study to minimize the influence of their fluctuations. We then consider the longitudinal data since the beginning of 6th year of the EDIC study. Unfortunately, possibly due to censoring, we see a high imbalance between events and censoring after year 12 of the EDIC study. Therefore, we limit our analysis to the 7 years following the first 5 years of the EDIC study. The final sample considered consists of 1292 individuals (47% females and 53% males) aged 28–54 years at the beginning of the 6th year of the EDIC study (our baseline). At the end of our observation interval, 70 of the 1292 individuals experienced event E.

3.1.2 Results on real data

Here, we describe the results of the application of the proposed approach, which is based on three different methods, to the above described database. We recall that the event E considered here consists of having one of the CVD events described in section 3.1.1, within a time interval of 7 years. For the first method, we excluded censored individuals. However, the fraction of censored individuals is small (3.5%). For the other two methods, we take advantage of having longitudinal data to compensate for the presence of censoring. In fact, the values of $\hat{p}_i$ introduced in the previous section are computed based

Table 2. Estimated values of the factor rate a for each factor, appearing in the linear model of equation (2) for 7-year risk probability of CVD in individuals with T1D.

Factor	Bayesian	Estimator variance	Unexplained fluctuations	Final estimate
SBP	0.46 (0.13)	0.37 (0.14)	0.39 (0.15)	0.41 (0.13)
LDL	0.10 (0.03)	0.11 (0.03)	0.11 (0.03)	0.11 (0.03)
TRG	0.052 (0.016)	0.047 (0.018)	0.049 (0.019)	0.049 (0.018)
HBA1C	0.43 (0.09)	0.36 (0.13)	0.37 (0.14)	0.39 (0.12)
HDL	0.49 (0.15)	0.61 (0.19)	0.62 (0.20)	0.58 (0.18)

Note: The values shown (standard deviation in parenthesis) represent the increase of the event probability (in percentage) corresponding to a unitary increase (decrease for HDL) of the factor. The factor units are reported in the second column of Table 1. For HBA1C, the parameter a is the one of the non-linear model in equation (3). For the non-Bayesian methods, the values shown are the arithmetic means of the values of the rates obtained for the combinations of the values of the number of sub-intervals and of the length of the interval I_0. The p-values associated to the rate estimates from the different combinations are all significant ($p < 0.05$). In the last column, the final estimates of the rates are shown, computed as averages of the values from the three methods. The corresponding standard deviations in parenthesis are also reported. CVD: cardiovascular disease; T1D: type 1 diabetes; SBP: systolic blood pressure; LDL: low density lipoprotein; HDL: high density lipoprotein; HBA1C: glycated hemoglobin level; TRG: triglyceride concentration.

on all individuals, without excluding those censored, by the values of the KM curve[17] at last time (end of year 7). For each of these two methods, the estimates of the rates a are obtained by averaging those obtained using a number of sub-intervals $k = 4, 5, 6$. The empirical ranges $[f_L, f_R]$ for the considered risk factors are reported in Table 1. When applying the method to find the range of analysis, we set $\delta = 1$ for the first four of them and $\delta = 0.1$ for HBA1C. We set the length of the interval I_0 equal to $\Delta = 10, 15, 20$ ($\Delta = 0.8, 1.0, 1.2$) for SBP, LDL, HDL and TRG (HBA1C).

In Table 2 are summarized the results of all five factors. For each factor, there is good agreement between the estimates of the rate a obtained from the three different methods. We have taken the arithmetic average of these three values of the rates obtaining the final ones as reported in the last column of Table 2. Since, from the simulation study, we find a high correlation between the estimates obtained from the three different methods, to estimate the standard deviation of the final estimates we proceed similarly as done separately for each non-Bayesian method. Indeed, we compute the standard deviation as an upper bound by considering each correlation equal to one. The resulting value of the standard deviation corresponds to the arithmetic average of those obtained from the three methods.

By applying the proposed detection method, a region with small factor values to be excluded from the analysis is detected for SBP ($f_A = 118$ mmHg), LDL ($f_A = 95$ mg/dL), TRG ($f_A = 56$ mg/dL) and HBA1C ($f_A = 8.1\%$). For HDL the region to be excluded is the one with factor values larger than $f_A = 51$ mg/dL. For each of the five factors, the p-value for the χ^2 test described in the previous section is very large. Therefore, data are compatible with the null hypothesis. We notice that, for each factor, the estimated value of f_A is close to the corresponding one of the distribution mode of that factor, shown in Table 1. However, when considering the first procedure, for SBP, LDL and TRG, we do not have evidence of fluctuations around zero of $p_R - p_L$ for a large set of sub-interval configurations whose common point f_0 is on the left of f_A, as said in the previous section. In fact, for SBP and TRG, we could not have f_0 far enough on the left of the mode. For LDL, although f_L allowed this, we have values of $p_R - p_L$ significantly smaller than zero for several sub-interval configurations. This suggests that for LDL we have and initial decreasing phase. Therefore, from all these considerations, we conclude that for SBP, LDL and TRG, we do not have evidence of a real biological activation phenomenon. The situation is different for HDL, where both conditions are met. In addition to the fulfillment of them, also the fluctuations of the difference considered in the second detection procedure are around zero for a large set of the sub-interval configuration on the right of f_A. Then, for HDL we have even more evidence than required that a real biological activation phenomenon exists. We notice that this is already reported in literature.[33] We only have partial evidence of such phenomenon for HBA1C. In fact, for this factor, the condition required on the fluctuations involved in the first procedure is satisfied, but the one on the location of f_A is violated for the second detection procedure.

We now illustrate separately the results for each of the five risk factors considered here. We highlight that, for each factor, integrating the estimated linear model over the whole empirical range, after weighting it by the probability density function $s(\cdot)$, we obtain a value very close to the corresponding risk estimated empirically from the whole sample.

3.1.3 Systolic blood pressure

Here, we estimate from the data a value of 118 mmHg for f_A. For the risk rate of increase a of the linear model, the value of the final estimate is 0.41%/mmHg (standard deviation 0.13%/mmHg). Therefore, for values of SBP > 118 and < 172 mmHg, an increase of 10 mmHg corresponds to an increase of the risk of about 4%.

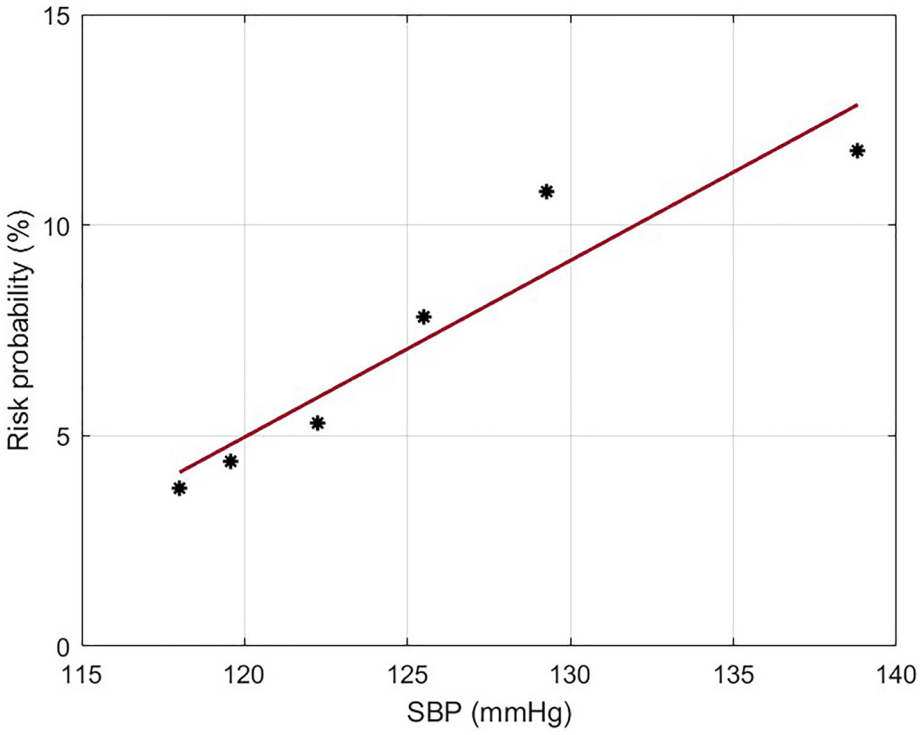

Figure 2. Estimated values of the 7-year risk versus the values of systolic blood pressure (SBP) for the optimal sub-interval configuration from the method based on unexplained fluctuations with $k = 5$. The length Δ of the interval I_0 is equal to 20 mmHg. The corresponding fitted theoretical linear model is superimposed to the empirical data. The first data point corresponds to the sub-interval I_0.

As an example, in Figure 2, we show the estimated value of the 7-year risk $\hat{p}_i$ for SBP as a function of the mean value $\bar{f}_i$ of that factor in the i-the element $\hat{I}_i$ of the optimal sub-interval configuration obtained by the method based on unexplained fluctuations with $k = 5$. The specific value of Δ used here and in the figures for the other factors are reported in their captions.

3.1.4 Low density lipoprotein cholesterol

For this risk factor, we have found $f_A = 95$ mg/dL. The value of the final estimate obtained for the risk rate of increase a of the linear model is 0.11%/(mg/dL) (standard deviation 0.03%/(mg/dL)). Therefore, for value of the LDL in $[95, 255]$ mg/dL, an increase of 10 mg/dL corresponds to an increase of the risk of about 1%. Figure 3 shows an example of the 7-year risk for the event E as a function of the mean values of LDL in the sub-intervals corresponding to the optimal division of the range of analysis using the estimator variance method with $k = 6$.

3.1.5 Triglyceride concentration

In this case, we have found $f_A = 56$ mg/dL. The value of the final estimate for the risk rate of increase of the linear model is equal to 0.049%/(mg/dL) (standard deviation 0.018%/(mg/dL)). Therefore, for values of TRG in $[56, 635]$ mg/dL, an increase of 100 mg/dL in TRG corresponds to an increase in risk of about 5%.

In Figure 4, we show the estimates of conditional probabilities $\hat{p}_i$ as a function of the corresponding averages of TRG values $\bar{f}_i$ in the sub-intervals I_i relative to the optimal division of the range of analysis according to the estimator variance method with $k = 5$.

3.1.6 High density lipoprotein cholesterol

We notice that for this risk factor, a real activation phenomenon is detected here, as already reported in literature.[33] Indeed, the variations of the risk appear only for values of the factor ≤ 51 mg/dL, this value is very close to the one obtained for females (50 mg/dL) in Alberti et al.[33] (40 mg/dL for males). Therefore, although this factor is known as a protecting one, increasing its value above 51 mg/dL does not make the risk of CVD to decrease. The value of the final estimate for the rate a of the linear model is 0.58%/(mg/dL) (standard deviation 0.18%/(mg/dL)). Therefore, for values of HDL in $[28, 51]$

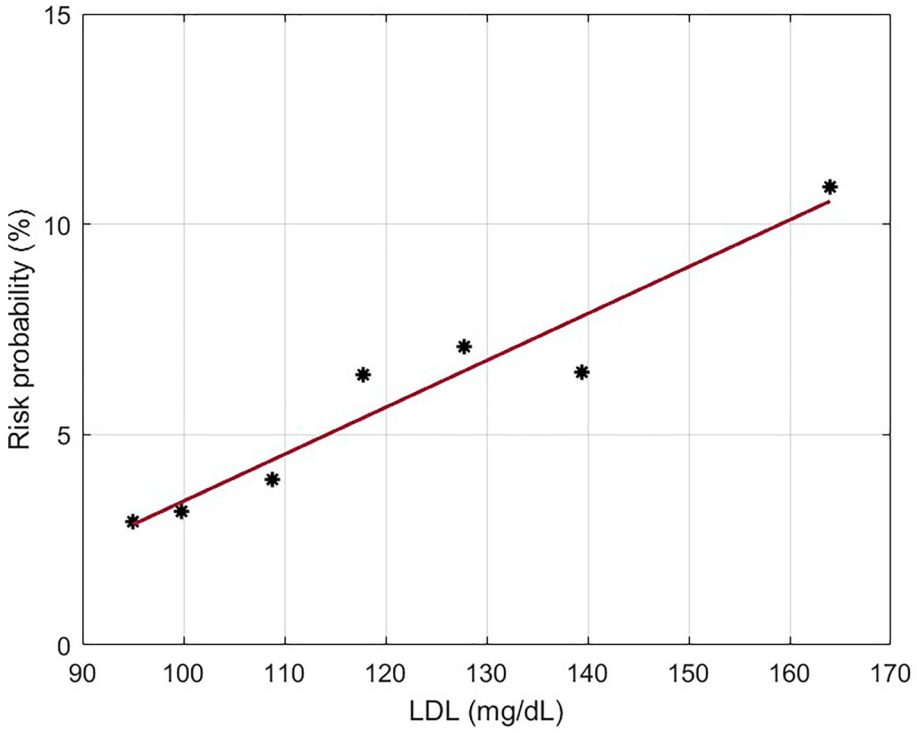

Figure 3. Estimated values of the 7-year risk versus the values of low density lipoprotein (LDL) for the optimal sub-interval configuration from the method based on the estimator variance with $k = 6$. We use $\Delta = 20$ mg/dL. The corresponding theoretical linear model is superimposed to the empirical data. The first data point corresponds to the sub-interval I_0.

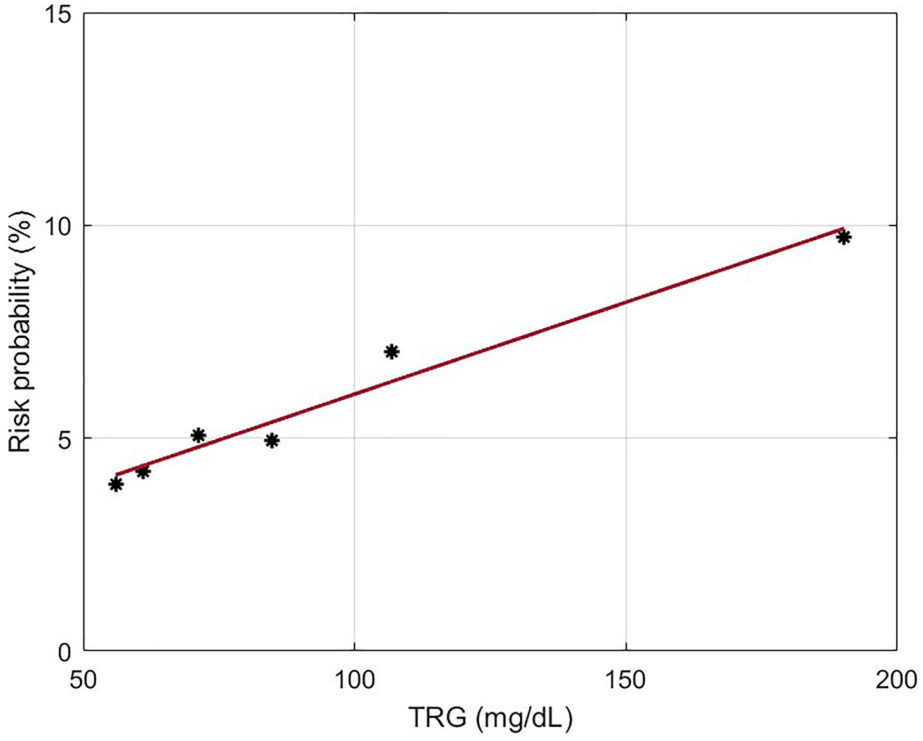

Figure 4. Estimated values of the 7-year risk versus the values of triglyceride concentration (TRG) for the optimal sub-interval configuration from the method based on the estimator variance with $k = 5$. Here, we use $\Delta = 15$ mg/dL. The corresponding theoretical linear model is superimposed to the empirical data. The first data point corresponds to the sub-interval I_0.

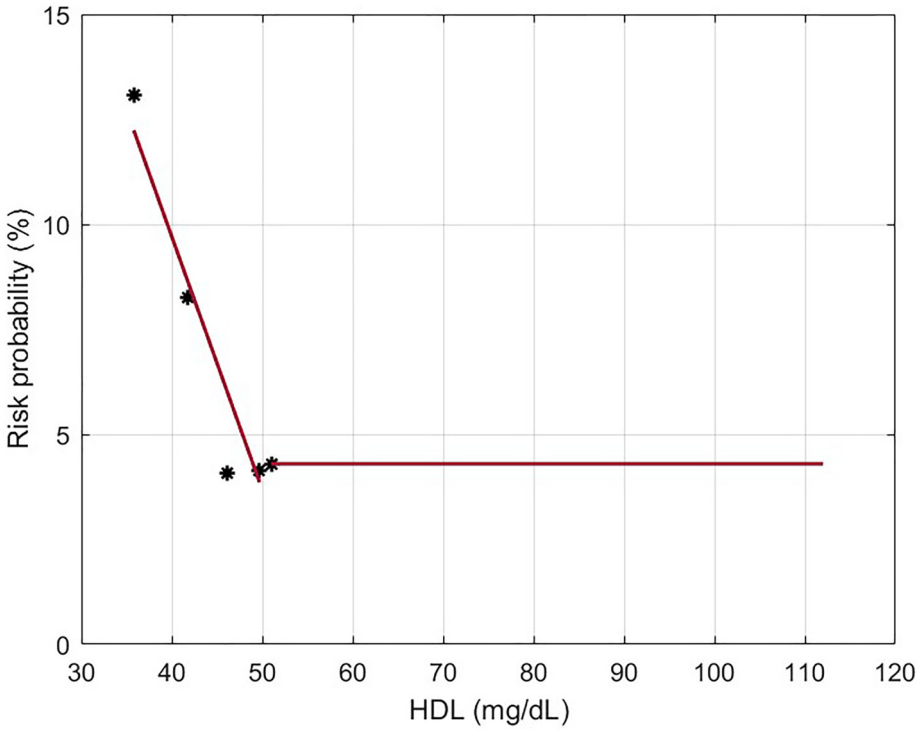

Figure 5. Estimated values of the 7-year risk versus the values of high density lipoprotein (HDL) for the optimal sub-interval configuration from the method based on the unexplained fluctuations with $k = 4$. The corresponding theoretical linear model is superimposed to the empirical data. Since for this factor we have a real activation phenomenon, the last data point does not corresponds to the sub-interval I_0, but to $[f_A, f_R] = [51, 112]$ mg/dL, and the horizontal line is relative to the risk of individuals with factor values belonging to that sub-interval.

mg/dL a decrease of 10 mg/dL corresponds to a risk increase of about 5.8%. Figure 5 shows the variation of the estimated value $\hat{p}_i$ for HDL as a function the mean value $\bar{f}_i$ for the optimal division of the range of analysis using the unexplained fluctuations method with $k = 4$.

3.1.7 Glycated hemoglobin level

The value found for f_A = is 8.1%. The value of the final estimate of the parameter a of the non-linear model is equal to 0.39/% (standard deviation 0.12/%). Therefore, for values of HBA1C in $[8.1, 13]$%, an increase of the factor by one unit, corresponds to a risk increase by a factor $e^{0.39} \simeq 1.48$. In Figure 6, are shown the empirical values of the 7-year risk as a function of the mean values of HBA1C for the optimal sub-interval configuration estimated with the unexplained fluctuations method with $k = 5$.

3.2 Application to simulated data

3.2.1 Data

To quantify the performances of the proposed methodology, we apply it several times to simulated datasets. To generate a realistic simulated dataset, we proceed as follows. First, we choose a parametric model to describe the probability of the event, conditioned on a single risk factor. We use here a linear model for all the factors, but HBA1C, for which we use an exponential model. The parameters of these models are the ones obtained applying the proposed methodology to the real dataset. We then consider the probability density function describing each factor in the real dataset, whose parameters are estimated from these data. For each factor, we use the corresponding theoretical distribution to draw a sample. Then, for each realization of the sample, we simulate the events using a Bernoulli distribution with probability equal to the value of the linear (or exponential) model corresponding to the factor value of that realization. Finally, we randomly censor some of the simulated observations, according to the proportion of censored individuals observed in the real dataset. We perform the estimation on a simulated dataset with a sample dimension $m = 1600$, which is close to that in the real dataset.

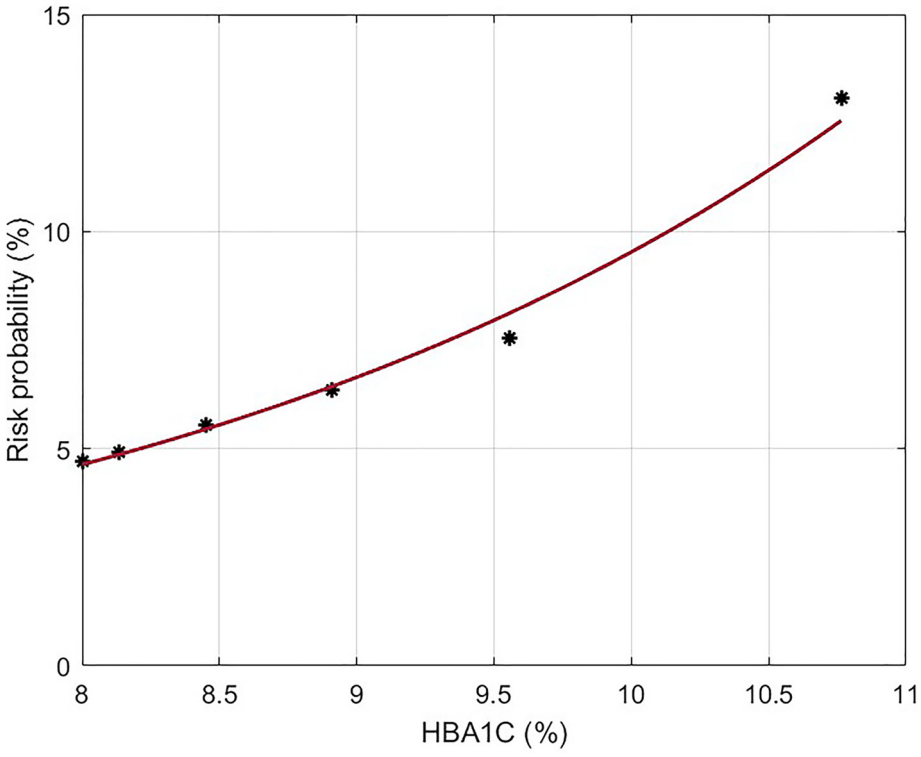

Figure 6. Estimated values of the 7-year risk versus the values of HBA1C for the optimal sub-interval configuration from the method based on the unexplained fluctuations with $k = 5$ and $\Delta = 1.2\%$. The corresponding theoretical non-linear model is superimposed to the empirical data. The first data point corresponds to the sub-interval I_0.

Table 3. Results of the application of the proposed methodology and of the standard maximum likelihood approach to simulated datasets, each with $m = 1600$ individuals.

Factor	Proposed method			Max Likelihood		
	discr	fluct	frac	discr	fluct	frac
SBP (0.41)	0.10	0.13	0.95	0.09	0.10	0.92
LDL (0.11)	0.02	0.03	0.97	0.02	0.02	0.89
TRG (0.049)	0.014	0.018	0.97	0.014	0.013	0.85
HBA1C (0.39)	0.08	0.11	0.98	0.07	0.06	0.87
HDL (0.58)	0.13	0.17	0.95	0.13	0.13	0.92

Note: The values reported are relative to 1000 independent random datasets. In the first column, we show the acronyms of the factors considered, together with their corresponding true values in parenthesis. For each method and factor, we report the mean absolute discrepancy between the estimated rate and the corresponding true value (*discr*). In addition, a measure of the rate estimate fluctuations is also shown (*fluct*). Finally, we include the fraction of replications for which the true value of the rate is contained in the interval centered at the estimate and with half width equal to two times the measure of fluctuation (*frac*). SBP: systolic blood pressure; LDL: low density lipoprotein; HDL: high density lipoprotein; HBA1C: glycated hemoglobin level; TRG: triglyceride concentration.

3.2.2 Results

In Table 3, we show results of the application of the proposed methodology to realistic simulated datasets. As a comparison, we also include the analogous results relative to the probit method in the version of Fisher where maximum likelihood inference is performed. To account for sampling variability, we simulate several independent random datasets. The number of simulations of the dataset is set such that the values of the quantities considered, described below, do not change when further increasing it.

For the two methods and each factor, we report three quantities. First, we compute the mean discrepancy (in absolute value) between the rate estimate and the corresponding true value. We also consider a measure of the fluctuations of the rate estimate. This is computed as the square root of the average value of the variances of the estimates relative to the

different replications. The estimate's variance of each dataset is calculated as described in the beginning of section 3.1.2. For each replication, as variance of the maximum likelihood estimate, we consider the lower bound $m/\mathcal{I}(\hat{a})$, where $\mathcal{I}(\hat{a})$ is the Fisher information relative to the slope estimate.[6] Finally, for each factor, we compute the fraction of replications such that the true value of the rate is contained into the interval centered at the estimate and with half width equal to two times the fluctuation measure.

4 Discussion

The methodology proposed in this paper has been applied to quantify the influence of some of the main factors, i.e. SBP, LDL, HDL, TRG and HBA1C, on the risk of CVD in individuals with T1D using a large database. The results obtained from the real data show that the theoretical model in equation (2) (equation (3) for HBA1C) performs well to describe the factors influence on the 7-year risk. Moreover, for each of the five factors considered, the values of the rates estimated by applying the three different methods considered are close to each other. The values of the coefficients of variation (standard deviation/average value) of the final rate estimates for the five factors considered are about 25–30%. When comparing these values, for example, with the ones reported by the STENO1 study,[13] we see that our method outperforms this last. In fact, the coefficients of variation for the continuous variables range from 35% to 41%.

We notice that using the simulated datasets, the three methods show only small differences of the performances. In particular, the difference in the absolute discrepancy between any pair of the three methods is at most 0.03. It happens exactly the same for the percentage of success in retrieving the true value. However, we notice that the Bayesian method presents absolute discrepancy equal or lower than those from the other two methods. A similar situation happens to the unexplained fluctuations method, when looking at the success percentages. Of course, the implementation effort and the computational cost of the three methods are quite different, with the variance based method being the easiest to handle. On the other hand, for the method based on unexplained fluctuations both the implementation effort and the computational cost are significantly larger. As pointed out, for the Bayesian method, there are different possibilities for the implementation involving different levels of implementation effort and computational cost. Furthermore, this last method is more flexible, allowing to include different types of *a priori* information. For example, it can be implemented within the empirical Bayes approach.[7]

As the simulation results demonstrate, the three methods are quite close to one another in terms of performances. Therefore, we decide to make them contributing equally to the arithmetic mean, which provides the final estimate of the rate. Another possibility could be to slightly increase of an equal amount the weights (of the convex combination) for the Bayesian and the unexplained fluctuations methods. Indeed, the first one shows a better behavior in terms of absolute mean discrepancy, while the other presents higher percentage of success in retrieving the true value.

We notice that the results from real data provide evidence for the existence of a real biological activation phenomenon for the protective factor HDL. Indeed here we can conclude that the risk does not change, reducing the value of HDL, until we reach 51 mg/dL, below which the risk increases linearly. We notice that such real phenomenon for HDL has already been reported in literature[33] with a value of the threshold close to the one obtained here. Unfortunately, we only have partial evidence of a similar phenomenon for HBA1C.

Although we apply the proposed methodology to data from individuals with T1D for the risk of CVD events, it can be used in general for risk quantification. Here we have provided methods details and have illustrated the results mostly in the case of a linear model $g(\cdot)$. However, the approach can also be used with minor modifications in case a non-linear function $g(\cdot)$ is adopted, and we have illustrated this for the case of HBA1C.

Realistic simulated data are used here to quantify the performances of the proposed methodology and to compare them to those from the probit method. We notice that the ability of the proposed methodology to correctly estimate the rate is good. Indeed, this is evident both from the discrepancy values and from the success percentage in retrieving the true value of the rate. Looking at the maximum likelihood approach, the results are qualitatively similar to those obtained with the proposed methodology. However, when comparing the fraction of replications for which the true value is successfully retrieved, the maximum likelihood values are lower for all factors. A possible explanation is that maximum likelihood approach underestimates the fluctuations of the estimated rates. This can be made evident comparing the columns *fluct* relative to the proposed methodology and the maximum likelihood approach. We finally notice that, for both method, when considering the mean discrepancy without absolute value, the results show a significant decrease respect the corresponding ones with absolute value, indicating that the estimates are unbiased.

We highlight that, the dose-effect curve estimated by the above methodology remains influenced by the effects on E of factors correlated to it, but not causally dependent on the factor f_c under study. We now see how we can cope, under some conditions, with this situation. Let us focus on the case of two non-binary risk factors f_c and f_t. For our application, the first factor could be for example the level of triglycerides and the second one the age. First of all, we resample the data available

based on the same (arbitrary) probability density function $q(\cdot)$ of the values of f_t conditioned to any value of f_c. Using the resampled data, we then compute the dose–response curve of f_c. Under the hypothesis that the two factors influence E in an independent way, the combined dose–response function is given by the product of two functions, dependent, one on f_c and the other one on f_t. Then, the probability of E in the population corresponding to this sub-sample is obtained by (double) integrating the product between these two functions, $q(\cdot)$ and the probability density function of f_c. Based on the above assumption, this double integral becomes proportional to the integral of the function on f_c weighted by the last density. Therefore, the dose–response curve estimated from the sub-sample is accounting only the effects of f_c. If f_c is a confounder, i.e. the event E is not at all influenced by it, but by another factor f_t correlated with f_c. Then, the above function dependent only on f_c, is constant. Therefore, the dose–response estimated from the sub-sample is flat, while this is not the case for the original sample. In general, we notice that in the linear case, under the same hypothesis, the estimate of the rate a from the sub-sample is not affected by the above mentioned effects. However, this is not in general true for the absolute level of the dose–response curve. To further study the situation, we could repeat several times the sub-sample procedure with different conditional distributions of the values of f_t given f_c, with different expected values. Then, by analyzing the set of dose–response curves obtained, we could extrapolate the one not influenced by f_t.

The proposed methodology implicitly takes into account associations between factors. In a different approach, qualitative information about these associations (causal relationships) can be provided by medical experts, based on the underlying biological mechanisms. Then, a specific probabilistic model, within the class of Bayesian Belief Networks (BBNs), can be built, whose parameters are estimated from data. We do not expect such method to be superior to standard risk calculators when performing risk prediction. However, after parameter estimation, the probability of the event E studied conditioned to any risk factors combination can be derived. In principle, such conditional probabilities could be directly estimated from data. However, due to the limited sample size in the database available, this limits the analysis to combinations of at most pairs of factors, e.g. SBP with smoking. Nevertheless, the last analysis could be useful to describe how the influence of a factor on the risk varies as a function of a second one, e.g. in an additive or a multiplicative way. Future research includes the development of a specific BBN model based on information from medical experts, its application to the same database, and the comparison of the results with those obtained by the methodology proposed in this paper.

5 Conclusions

In this paper, we propose a new methodology to quantify the influence of single risk factors on the probability of adverse event occurrence. We consider here both the case of linear dependence between variable and risk and the non-linear one. The methodology is based on the combination of three different methods. For each factor, the estimates of the rates obtained from the three methods are close to each other. Nevertheless, we provide the final rate estimate combining them to increase the results reliability. The results obtained from a large realistic simulation study show that the proposed methodology outperforms the standard method used to deal with this type of models. It can be used in general, although it is applied here for CVD risk in people with T1D. In addition, we propose the combination of two different procedures to possibly reduce the range of the factor values where the analysis is performed. This is done to cope with the situations where data do not provide evidence of risk changes in a most left or most right part of the range values of the factor. In case some additional conditions are verified, this allows to detect the presence of a real biological phenomenon of activation. This occurs here for HDL, for which we find a value above which the risk does not decrease when increasing the value of this factor. This phenomenon has already been described and the threshold value obtained from our analysis is close to those previously reported in the literature.

Acknowledgments

The authors are thankful to Fred Godtliebsen, Conceição Granja and the anonymous reviewers for their useful comments and suggestions.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017385 (Art. 29.1 GA).

Data acknowledgments

The DCCT and its follow-up the EDIC study were conducted by the DCCT/EDIC Research Group and supported by National Institute of Health grants and contracts and by the General Clinical Research Center Program, NCRR. The data from the DCCT/EDIC study were supplied by NIDDK Central Repository. This manuscript was not prepared under the auspices of the DCCT/EDIC study and does not represent analyses or conclusions of the DCCT/EDIC study group, NIDDK Central Repository, or NIH.

ORCID iDs

Ornella Moro https://orcid.org/0000-0002-3607-1797
Marit B Veierød https://orcid.org/0000-0002-2083-2758

References

1. Wilson PWF, et al. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**: 1837–1847.
2. Yusuf S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study. *The Lancet* 2004; **364**: 937–952.
3. Brown CC. The statistical analysis of dose-effect relationships. In: G.C. Butler. *Principle of Ecotoxicology*. John Wiley & sons, 1978.
4. Bliss CI. The method of probits. *Science* 1934; **79**: 38–39.
5. Fisher RA. Appendix: the case of zero survivors. C I bliss "the calculation of the dosage-mortality curve". *Ann Appl Biol* 1935; **221**: 134–167.
6. Lehmann EL. *Elements of large-sample theory*. New York, NY: Springer, 1999.
7. Carlin BP and Louis TA. *Bayesian methods for data analysis*. Boca Raton, FL: Chapman and Hall/CRC Press, 2008.
8. Kirkpatrick S, Gelatt CD and Vecchi MP. Optimization by simulated annealing. *Science* 1983; **220**: 671–680.
9. Geman S and Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 1984; **PAMI-6**: 721–741. DOI: 10.1109/TPAMI.1984.4767596.
10. Cox DR and Oakes D. *Analysis of survival data*. Boca Raton, FL: Chapman and Hall/CRC Press, 2018.
11. Selmer R, et al. Norrisk 2: a norwegian risk model for acute cerebral stroke and myocardial infarction. *Eur J Prev Cardiol* 2017; **24**: 773–782.
12. SCORE2 working group and ESC Cardiovascular risk collaboration. Score2 risk prediction algorithms. new models to estimate 10-year risk of cardiovascular disease in europe. *Eur Heart J* 2021; 42: 2439–2454.
13. Vistisen D, et al. Prediction of first cardiovascular disease event in type 1 diabetes mellitus. *Circulation* 2016; **133**: 1058–1066.
14. Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC). https://repository.niddk.nih.gov/studies/edic/?query=cardiovascular (accessed November 2024).
15. Diabetes Control Complications Trial Research Group. Design and methodologic considerations for the feasibility phase. The DCCT research group. *Diabetes* 1986; **35**: 530–545.
16. Epidemiology of Diabetes Interventions Complications (EDIC) Research Group. Design, implementation, and preliminary results of a long-term follow-up of the Diabetes Control and Complications Trial cohort. *Diabetes Care* 1999; **22**: 99. DOI: 10.2337/diacare.22.1.99.
17. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; **53**: 457–481.
18. Ambrose JA and Barua RS. The pathophysiology of cigarette smoking and cardiovascular disease: an update. *J Am Coll Cardiol* 2004; **43**: 1731–1737.
19. Benowitz NL. The role of nicotine in smoking-related cardiovascular disease. *Prev Med* 1997; **26**: 412–417.
20. Smith CJ and Fischer TH. Particulate and vapor phase constituents of cigarette mainstream smoke and risk of myocardial infarction. *Atherosclerosis* 2001; **158**: 257–267.
21. Aberra T, et al. The association between triglycerides and incident cardiovascular disease: what is "optimal"? *J Clin Lipidol* 2020; **14**: 438–447.e3.
22. Gilks WR, Richardson S and Spiegelhalter D. *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman and Hall/CRC Press, 1995.
23. Metropolis N, et al. Equation of state calculations by fast computing machines. *J Chem Phys* 1953; **21**: 1087–1092.
24. Madras NN. *Lectures on Monte Carlo methods*, volume 16. American Mathematical Society, 2002, pp. 90 – 92.
25. Nadaraya EA. On estimating regression. *Theory Probab Appl* 1964; **9**: 141–142.
26. Watson GS. Smooth regression analysis. *Sankhyā: Indian J Stat Ser A* 1964; **26**: 359–372.
27. Wilcoxon F. Individual comparisons by ranking methods. In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 196–202.
28. Smith C. Chi-squared tests with small numbers. *Ann Hum Genet* 1986; **50**: 163–167.
29. Mortensen MB and Nordestgaard BG. Elevated LDL cholesterol and increased risk of myocardial infarction and atherosclerotic cardiovascular disease in individuals aged 70–100 years: A contemporary primary prevention cohort. *The Lancet* 2020; **396**: 1644–1652.
30. Colom C, Rull A, Sanchez-Quesada JL et al. Cardiovascular disease in type 1 diabetes mellitus: Epidemiology and management of cardiovascular risk. *J Clin Med* 2021; **10**: 1798. https://www.mdpi.com/2077-0383/10/8/1798.

31. Htay T, et al. Mortality and cardiovascular disease in type 1 and type 2 diabetes. *Curr Cardiol Rep* 2019; **21**: 1–7.
32. Rawshani A, et al. Range of risk factor levels: control, mortality, and cardiovascular outcomes in type 1 diabetes mellitus. *Circulation* 2017; **135**: 1522–1531.
33. Alberti KGMM, et al. Harmonizing the metabolic syndrome. *Circulation* 2009; **120**: 1640–1645.

Appendix A

A problem with existing risk calculators, which are based on standard approach, i.e. CPH model, arises. In fact, let us consider the case when a binary random variable C (representing the event E) is influenced by other two (factors), A and B, also binary and independent to each other. Assuming the Cox model, from the above formulas, we can express the hazard, i.e. the probability that the output event happens (C=1) in a small time interval δt after the time $t = 0$, given the values of the two random factors A and B and that the event did not happen before time $t = 0$, in the simplest case of an hazard constant along time, as

$$p(C = 1, t < \delta t | A = a, B = b, C = 0\,\forall t' \leq 0) = \delta t\, e^{\beta_a a + \beta_b b}$$

where $a, b = 1, 0$. As commonly done when adopting such model, to quantify the influence of each of the two factors A and B on the output event C, we can consider the ratio between the above expression corresponding to the values 1 and 0:

$$\frac{p(C = 1, t < \delta t | A = 1, B = b, C = 0\,\forall t' \leq 0)}{p(C = 1, t < \delta t | A = 0, B = b, C = 0\,\forall t' \leq 0)} = e^{\beta_a}$$
$$\frac{p(C = 1, t < \delta t | A = a, B = 1, C = 0\,\forall t' \leq 0)}{p(C = 1, t < \delta t | A = a, B = 0, C = 0\,\forall t' \leq 0)} = e^{\beta_b}$$

Therefore, if $\beta_b > \beta_a$, it follows that $e^{\beta_b} > e^{\beta_a}$, so that the influence of B on the risk of event C is larger that the one corresponding to A.

However, in the same situation, if we consider a random dependence between the two factors, the actual influence on the output event of each of them separately may be quite different. Let us assume, for example, that the random factor B depends on the other one A, as shown by the directed acyclic graph in Figure A1. Then, we can write

$$\begin{aligned} p(C = 1, t < \delta t | A = a, C = 0\,\forall t' \leq 0) &= \frac{\delta t}{p(A = a)} \sum_b e^{\beta_a a + \beta_b b} p(B = b | A = a) p(A = a) \\ &= \delta t e^{\beta_a a} \sum_b e^{\beta_b b} p(B = b | A = a) \end{aligned}$$

where $a = 1, 0$.

Taking the ratio between the expressions above for $a = 1$ and $a = 0$, we obtain

$$\begin{aligned} \frac{p(C = 1, t < \delta t | A = 1, C = 0\,\forall t' \leq 0)}{p(C = 1, t < \delta t | A = 0, C = 0\,\forall t' \leq 0)} &= \frac{e^{\beta_a}(e^{\beta_b} p(B = 1 | A = 1) + p(B = 0 | A = 1))}{e^{\beta_b} p(B = 1 | A = 0) + p(B = 0 | A = 0)} \\ &= \frac{e^{\beta_a}(1 + (e^{\beta_b} - 1) p(B = 1 | A = 1))}{1 + (e^{\beta_b} - 1) p(B = 1 | A = 0)} \end{aligned}$$

In the limit $p(B = 1|A = 1) \to 1$ and $p(B = 1|A = 0) \to 0$, the last expression of the above ratio tends to $e^{\beta_a + \beta_b}$.

Let us now focus on the influence of factor B:

$$\begin{aligned} p(C = 1, t < \delta t | B = b, C = 0\,\forall t' \leq 0) &= \frac{\delta t}{p(B = b)} e^{\beta_b b} \sum_a e^{\beta_a a} p(B = b | A = a) p(A = a) \\ &= \frac{\delta t e^{\beta_b b}(e^{\beta_a} p(B = b | A = 1) p(A = 1) + p(B = b | A = 0) p(A = 0))}{p(B = b | A = 1) p(A = 1) + p(B = b | A = 0) p(A = 0)} \end{aligned}$$

where $b = 1, 0$.

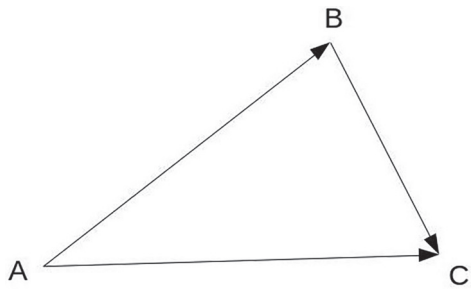

Figure A1. Directed acyclic graph for an event C dependent on two factors A and B.

In the limit $p(A = 1)/p(B = 1|A = 0) \to 0$, we have $p(C = 1, t < \delta t|B = b, C = 0\,\forall t' \leq 0) \to \delta t e^{\beta_b b}$, and the ratio between its values corresponding to $b = 1$ and $b = 0$ tends to

$$\frac{p(C = 1, t < \delta t|B = 1, C = 0\,\forall t' \leq 0)}{p(C = 1, t < \delta t|B = 0, C = 0\,\forall t' \leq 0)} \to e^{\beta_b}$$

Therefore, since $\beta_a > 0$, it follows that $e^{\beta_a+\beta_b} > e^{\beta_b}$, so that the influence of factor A on the output event C is larger than the one of B. We notice that the conclusion derived by using standard approach is the opposite and therefore it is wrong.

Appendix B

Let us denote by $g(\cdot)$ a non decreasing function which represents the value of the risk probability of a binary event E conditioned that the value of a given risk factor is f, i.e. $g(f) := \mathcal{P}(E = 1 \mid f)$. Let us assume that $g(\cdot)$ is a continuous function qualitatively described in Figure B1, with a jump in the derivative at f_A. In the first part ($f < f_A$), the function is constant, while it grows linearly with a positive slope in the second part. Now, we consider a small sub-interval $J_0 := [f_0, f_0 + \Delta]$ lying on the right on the activation point f_A ($f_A < f_0$). We now prove that the difference of the expected values of the factor, conditioned on the two possible values of the event E, $\mu_E - \mu_{\overline{E}}$, is positive

$$\mu_E - \mu_{\overline{E}} = \mathbb{E}(f|E, f \in J_0) - \mathbb{E}(f|\overline{E}, f \in J_0) > 0.$$

Indeed, we can rewrite the difference as follows:

$$\begin{aligned}
&\frac{\int_{J_0} fg(f)s(f)df}{\int_{J_0} g(f)s(f)df} - \frac{\int_{J_0} f(1 - g(f))s(f)df}{\int_{J_0} (1 - g(f))s(f)df} \\
&= \frac{\int_{J_0} fg(f)s(f)df}{\int_{J_0} g(f)s(f)df} - \frac{\mathcal{M}_1(J_0) - \int_{J_0} fg(f)s(f)df}{\mathcal{M}_0(J_0) - \int_{J_0} g(f)s(f)df} \\
&= \frac{\int_{J_0} f(a \cdot f + b)s(f)df}{\int_{J_0} (a \cdot f + b)s(f)df} - \frac{\mathcal{M}_1(J_0) - \int_{J_0} f(a \cdot f + b)s(f)df}{\mathcal{M}_0(J_0) - \int_{J_0} (a \cdot f + b)s(f)df}
\end{aligned}$$

where we denote with $\mathcal{M}_0(J_0) = \int_{J_0} s(f)df$, and with $\mathcal{M}_1(J_0) = \int_{J_0} fs(f)df$. Let us denote the numerator and the denominator in the first term of the last expression as $\mathcal{N}$, and $\mathcal{D}$, respectively. We can rewrite the above difference as

$$\frac{\mathcal{N}}{\mathcal{D}} - \frac{\mathcal{M}_1(J_0) - \mathcal{N}}{\mathcal{M}_0(J_0) - \mathcal{D}} = \frac{\mathcal{N} \cdot \mathcal{M}_0(J_0) - \mathcal{N}\mathcal{D} - \mathcal{D} \cdot \mathcal{M}_1(J_0) + \mathcal{N}\mathcal{D}}{\mathcal{D} \cdot (\mathcal{M}_0(J_0) - \mathcal{D})}$$

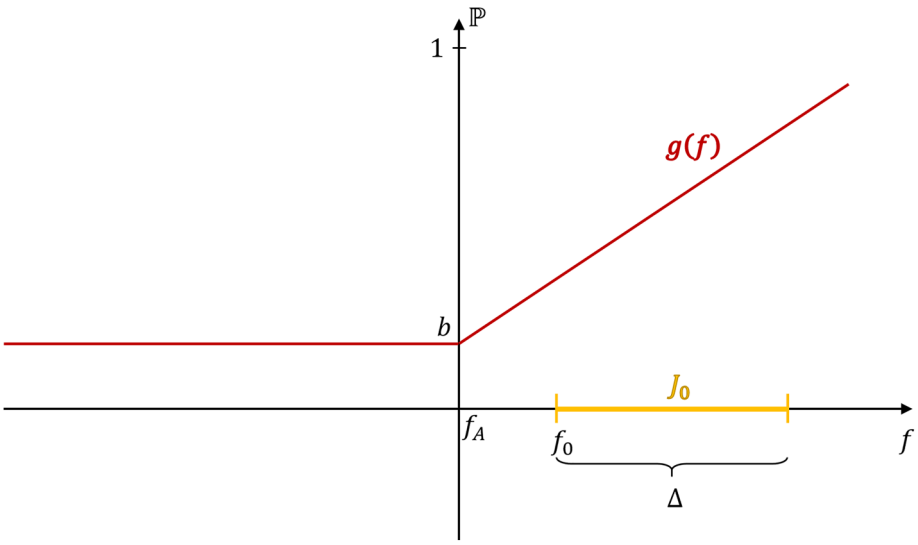

Figure B1. Profile of the $g(f)$ function.

From this we get

$$\frac{a\mathcal{M}_2(J_0)\mathcal{M}_0(J_0) + b\mathcal{M}_1(J_0)\mathcal{M}_0(J_0) - a(\mathcal{M}_1(J_0))^2 - b\mathcal{M}_1(J_0)\mathcal{M}_0(J_0)}{(a\mathcal{M}_1(J_0) + b\mathcal{M}_0(J_0))(\mathcal{M}_0(J_0) - a\mathcal{M}_1(J_0) - b\mathcal{M}_0(J_0))}$$
$$= \frac{a(\mathcal{M}_2(J_0) \cdot \mathcal{M}_0(J_0) - (\mathcal{M}_1(J_0))^2)}{(a\mathcal{M}_1(J_0) + b\mathcal{M}_0(J_0))(\mathcal{M}_0(J_0) - a\mathcal{M}_1(J_0) - b\mathcal{M}_0(J_0))} \tag{B1}$$

where we denote with $\mathcal{M}_2(J_0) = \int_{J_0} f^2 s(f)df$. Now, we can rewrite the last expression using conditional moments $\widetilde{\mathcal{M}}_i(J_0)$ $(i = 1, 2)$ of the factor f on the interval J_0, namely

$$\widetilde{\mathcal{M}}_i(J_0) = \frac{\mathcal{M}_i(J_0)}{\mathcal{M}_0(J_0)} = \int_{J_0} f^i \tilde{s}(f)df \tag{B2}$$

where we denote with $\tilde{s}(f) = \dfrac{s(f)}{\displaystyle\int_{J_0} s(f)df}$. This way, the last expression in equation (B1) becomes

$$\frac{a(\widetilde{\mathcal{M}}_2(J_0) - (\widetilde{\mathcal{M}}_1(J_0))^2)}{(a\widetilde{\mathcal{M}}_1(J_0) + b)(1 - a\widetilde{\mathcal{M}}_1(J_0) - b)} \tag{B3}$$

Since $\widetilde{\mathcal{M}}_1(J_0)$ is equal to the conditional expected value $\int_{J_0} f\tilde{s}(f)df$ of f on the interval J_0, we can rewrite the numerator of the last expression of $\mu_E - \mu_{\overline{E}}$ as follows

$$a(\widetilde{\mathcal{M}}_2(J_0) - (\widetilde{\mathcal{M}}_1(J_0))^2) = a \cdot \int_{J_0} (f - \widetilde{\mathcal{M}}_1(J_0))^2 \tilde{s}(f)df$$

The last integral is the conditional variance of the factor f on the interval J_0 and therefore it is always greater than zero. Since $a > 0$ and the denominator of $\mu_E - \mu_{\overline{E}}$ is positive from the beginning, this implies that also $\mu_E - \mu_{\overline{E}}$ itself is larger than zero.