Supplemental Digital Content

Manuscript: “Joint effects of intensity and duration of cigarette smoking on the risk of head and neck cancer: A bivariate spline model approach”

eAppendix

Selection of subjects (extended)

We provide some details on the extra checks on exposure definitions carried out to make our information more coherent and reliable (Figure 1).

We double-checked information on duration of cigarette smoking, starting from information on age, age at start, and age at quitting cigarette smoking. If duration was missing (464 subjects) and we had a coherent information on age, age at start, and age at quitting cigarette smoking (433 subjects), we imputed the missing values with the difference between age at stop smoking and age at start smoking, when these variables were available (37 subjects imputed and inserted in the statistical analysis). When duration was greater than age (20 subjects) and we had a reliable information on age, age at start, and age at stop of cigarette smoking (15 subjects), we similarly imputed duration with the difference between age at stop smoking and age at start smoking (5 subjects imputed and inserted in the statistical analysis).

Likewise, we double-checked information on intensity of cigarette smoking using information on pack-years of cigarette smoking. We excluded 1501 subjects with missing information on both intensity and pack-years of cigarette smoking. In addition, we

excluded 21 subjects with an implausible value of intensity of cigarette smoking (99 cigarettes per day).

Finally, we evaluated consistency of information of cigarette intensity and cigarette duration looking at “ever cigarettes” and “cigarette smoking status” variables. As there was perfect agreement between “Never” categories of variables “ever cigarettes” and “cigarette smoking status” and “Ever” category of “ever cigarettes” and “Former” and “Current” categories of “cigarette smoking status”, we referred to “ever cigarettes” in the following for simplicity. When “ever cigarettes” was equal to “Never”, we forced duration and intensity to be equal to zero (41 subjects). Similarly, if “ever cigarettes” variable was equal to “Ever”, we forced duration and intensity to be equal to one, which is the minimum value for a patient who smokes (5 subjects). When “ever cigarettes” showed a missing value (13 subjects), we were able to impute consistently “ever cigarettes” and “smoking status” using complete information from duration of cigarette smoking, age at start/stop of cigarette smoking, and cigarette intensity for 11 subjects.

Statistical analysis (extended)

The dose–response relationship between cancers of the oral cavity and pharynx (OCP) or larynx and the joint exposure to cigarette-smoking intensity and duration was investigated through bivariate regression spline models [1]. Briefly, we assumed a generalized semi-parametric model where the two exposures were entered as a joined piecewise polynomial of a linear degree, with constraints for continuity at each join point (called knot), together with potential confounders expressed in a parametric form:

$$\mathbf{logit}(\boldsymbol{\pi}) = \mathbf{C}\boldsymbol{\beta} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z})$$

where $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z})$ is defined as a bivariate regression spline with truncated linear basis:

$$f(x,z) = \delta_1 x + \delta_2 z + \delta_3 xz + \sum_{i=1}^{K_x} \gamma_i^{(x)} \left(x - \xi_i^{(x)}\right)_+ + \sum_{j=1}^{K_z} \gamma_j^{(z)} \left(z - \xi_j^{(z)}\right)_+$$

$$+ \sum_{i=1}^{K_x} \gamma_i^{(xz)} \left(x - \xi_i^{(x)}\right)_+ z + \sum_{j=1}^{K_z} \gamma_j^{(zx)} \left(z - \xi_j^{(z)}\right)_+ x + \sum_{i=1}^{K_x} \sum_{j=1}^{K_z} \gamma_{i,j}^{(xz)} \left(x - \xi_i^{(x)}\right)_+ \left(z - \xi_j^{(z)}\right)_+$$

x and z represent the two continuous exposures, $\xi^{(x)}$ and $\xi^{(z)}$ represent the knot positions, K_x and K_z represent the number of knots, and C represents the set of confounding factors given by {age, sex, race, study, education, alcohol drinking status, alcohol drinking intensity (number of drinks per day), and alcohol drinking duration} (see Table 1 for a complete list of the covariate categories used) [2].

We further assumed that the knot locations for any of the two exposures were unknown parameters to be estimated given a fixed combination of knot numbers up to a maximum of 2 [say, for intensity and duration respectively, (0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2)], in the absence of any evidence in favor of more than 2 knots.

For each cancer site, smoking status stratum, and combination of number of knots, we fitted the previous model using the Bayesian approach, as it allowed to: 1. identifying optimal knot locations starting from the overall set of confounding variables; 2. jointly estimating optimal knot locations and regression parameters; 3. formalizing constraints

on the knot locations through the definition of suitable and weakly informative prior distributions. In detail, we chose the prior distributions to express vague knowledge on plausible values of the parameters: for the knot locations, we assumed uniformly distributed priors on the range of the linked risk factor, subject to order constraints; for the regression parameters, we assumed Student-t distributions with 3 degrees of freedom and scale parameter equal to 10 for the intercept and to 2.5 for all the others regression parameters [3,4].

Posterior inference was then obtained combining information from the prior distributions and the likelihood function within the described Bayesian model via Markov Chain Monte Carlo (MCMC) simulation:

$$\boldsymbol{\pi}(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\xi} | \boldsymbol{y}, \boldsymbol{C}, \boldsymbol{x}, \boldsymbol{z}) \propto \boldsymbol{L}(\boldsymbol{y} | \boldsymbol{C}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \boldsymbol{\pi}(\boldsymbol{\beta}) \boldsymbol{\pi}(\boldsymbol{\delta}) \boldsymbol{\pi}(\boldsymbol{\gamma}) \boldsymbol{\pi}(\boldsymbol{\xi}),$$

where $\boldsymbol{L}(\boldsymbol{y} | \boldsymbol{C}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ is the likelihood of the logistic regression model and $\boldsymbol{\pi}(\boldsymbol{\beta})$, $\boldsymbol{\pi}(\boldsymbol{\delta}), \boldsymbol{\pi}(\boldsymbol{\gamma}), \mathbf{and}\ \boldsymbol{\pi}(\boldsymbol{\xi})$ represent the product of the independent prior distributions of the parameters. In detail, we simulated our model using the NUTS (No-U-Turn Sampler) algorithm [5], which is an MCMC algorithm based on the adaptive Hamiltonian Monte Carlo**.** We ran 4 chains of 2000 iterations each, with a warm-up period of 1000 iterations**.** Each of 4 chains was initialized in the following way: for the knot locations, we started the chains sampling a value at random from each exposure's range; for the regression parameters, we started every chain from zero.

A joint posterior distribution on knot locations and regression parameters was separately simulated for each cancer site, smoking status stratum, and combination of number of knots; it was based on the 4000 iterations (1000 iterations times 4 chains) of the corresponding sampling step. Diagnostics criteria, including trace plots of the marginal chains, $\hat{R}$ of single parameters ($1<\hat{R}<1.05$), divergent transitions (not present), and energy plots (histograms completely overlapped), were satisfied for most models and

reassured that the chains converged and the parameter space was fully explored for any parameter [6,7].

For each cancer site and stratum, we then chose the best model as the one that minimized the Watanabe-Akaike Information Criterion (WAIC) [8,9] among the convergent models with varying combinations of knots numbers that show plausible knot locations (knot location < 95^{th} percentile for either exposures).

Once the optimal combination of knot locations was identified for each cancer site and stratum, we calculated the corresponding ORs of cancers of the OCP and larynx, together with the 95% credible intervals (CIs), from the marginal posterior distribution of the parameters for the two exposures.

Calculations were carried out using the open-source Stan program [10], the open-source R program [11,12], its packages “rstan” [10], “loo” [13], “bayesplot” [14], and a specialized code that implemented bivariate spline models.

References

[1] Ruppert D, Wand M, Carroll R. Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. 2003.

[2] Dal Maso L, Torelli N, Biancotto E, et al. Combined effect of tobacco smoking and alcohol drinking in the risk of head and neck cancers: A re-analysis of case-control studies using bi-dimensional spline models. Eur J Epidemiol 2016; 31:385-393.

[3] Gelman A, Jakulin A, Pittau MG. Su YS. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 2008; 2:1360-1383.

[4] Ghosh J, Li Y, Mitra R. On the use of Cauchy prior distributions for Bayesian logistic regression. Bayesian Anal 2018; 13:359-383.

[5] Hoffman MD, Gelman A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 2014; 15:1593-1623.

[6] Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. 2017; eprint arXiv preprint arXiv:1701.02434.

[7] Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. Visualization in Bayesian workflow. J R Statist Soc A 2019; 182, Part 1:1-14.

[8] Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 2010; 11:3571-3594.

[9] Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Stat Comput 2014: 24(6), 997-1016.

[10] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. 2017; http://mc-stan.org.

[11] Ihaka R, Gentleman R. R: A language for data analysis and graphics. J Comput Graph Stat 1996; 5:299-314.

[12] **R** Development Core Team. **R**: A language and environment for statistical computing. **R** Foundation for Statistical Computing. Vienna: Austria. 2018.

[13] Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 2017; 27:1413-1432.

[14] Gabry, J. Bayesplot: Plotting for Bayesian models. 2017; R package version, 1(0).

eTable 1 - Characteristics of the individual studies from the International Head and Neck Cancer Epidemiology (INHANCE) consortium used in the current analysis

Study name, location	Case source	Age eligiblity	Case participation rate, %	Control source	Control participation rate, %	Matched factors	Recruitment period
Milan (1984-1989), Italy	Hospital	<80	95[a]	Hospital - unhealthy	95[a]	--	1984-1989
Aviano, Italy	Hospital	>18	>95[a]	Hospital - unhealthy	95[a]	--	1987-1992
Italy Multicenter	Hospital	18-80	>95	Hospital - unhealthy	>95	--	1990-1999
Switzerland	Hospital	<80	>95	Hospital - unhealthy	>95	--	1991-1997
Central Europe	Hospital	≥15	96	Hospital - unhealthy	97	Age, sex, ethnicity, city	1998-2003
New York, NY, USA (multicenter)	Hospital	21-80	91	Hospital- unhealthy	97	Age, sex, hospital, year of interview	1981-1990
Seattle (1985-1995), WA, USA	Cancer registry	18-65	54.4,63.3[b]	Random digit dialing	63.0,60.9[b]	Age, sex	1985-1995
Iowa, IA, USA	Hospital	>18	87	Hospital - unhealthy	92	Age, sex	1993-2006
North Carolina (1994-1997), NC, USA	Hospital	>17	88	Hospital - unhealthy	86	Age, sex	1996-1997
Tampa, FL, USA	Hospital	≥18	98	Hospital - noncancer	90	Age, sex, ethnicity	1994-2003
Los Angeles, CA, USA	Cancer registry	18-65	49	Neighborhood	68	Age, sex, neighborhood	1999-2004
Houston, TX, USA	Hospital	≥18	95	Hospital visitors	>80	Age, sex, ethnicity	2001-2006
Puerto Rico	Cancer registry	21-79	71	Residential records (healthy population)	83	Age, sex	1992-1995
Latin America	Hospital	15-79	95	Hospital - unhealthy	86	Age, sex, ethnicity, city	2000-2003
International Multicenter, IARC	Hospital	NA	88.7	Hospital / Community	87,3	Age, sex, center	1992-1997
Boston, MA, USA	Hospital	≥18	88.7	Residential records	48,7	Age, sex, neighborhood	1999-2003
Rome, Italy	Hospital	>18	98	Hospital - unhealthy	94	no matching	2002-2007
US Multicenter, USA	Cancer registry	18-79	75	Random digit dialing and health care rosters	76	Age, sex, ethnicity	1983-1984
Sao Paulo, Brazil	Hospital	NA	--	Hospital - unhealthy	--	Age, sex, city of residence, hospital	2002-2007
New York (MSKCC), NY, USA	Hospital	NA	--	Blood donors	--	Age, sex	1992-1994
Seattle-Leo, WA, USA	Cancer registry	20-74	81	Random digit dialing	75	Age, sex	1983-1987
Western Europe (ARCAGE)	Hospital	NA	82	Hospital - unhealthy (population based for UK centers)	68	Age, sex, ethnicity, city	2000-2005

Germany – Saarland	Hospital	50-75	94	Health examination	--	Age, sex	2001-2003
Germany – Heidelberg	Hospital	<80	96	Population registries	62,4	Age, sex, residence	1998-2000
Japan (2001-2005)	Cancer Hospital	20-79	97	Hospital - unhealthy	97	Age, sex	2001-2005
North Carolina (2002-2006), NC, USA	Cancer registry	20-80	82	DMV files	61	Age, sex, ethnicity	2002-2006
Paris (1989-1991), France	Hospital	NA	80	Hospital	86	Age, men only	1989-1991
HOTSPOT, USA	Hospital	≥18	>85	Hospital (benign conditions)	>80	Age, sex, race	2009-present
Buffalo, NY, USA	Hospital	NA	~50	Hospital	~50	Age, sex	1982-1998
Paris (2001-2007), France	Cancer registry	<=75	82.5	Random digit dialing	80.6	Age, sex, region	2001-2007
Baltimore, MD, USA	Hospital	NA	100	Hospital - benign conditions	70	Age, sex, HPV status	2000-2005
Beijing, China	Hospital	18-80	100	Hospital	100	Age, sex	1988-1989
Milan (2006-2009), Italy	Hospital	18-80	>95	Hospital	>95	--	2006-2009

ABBREVIATIONS: ARCAGE: Alcohol-Related Cancers And Genetic susceptibility in Europe; DMV: Department of Motor Vehicles; HPV: Human Papilloma Virus; IARC: International Agency for Research on Cancer; MSKCC: Memorial Sloan Kettering Cancer Center; NA: Not Available.

[a] Participation rate was not formally assessed, estimated response rate reported.

[b] Two response rates are reported because data were collected in two population-based case-control studies, the first from 1985 to 1989 among men and the second from 1990 to 1995 among men and women.

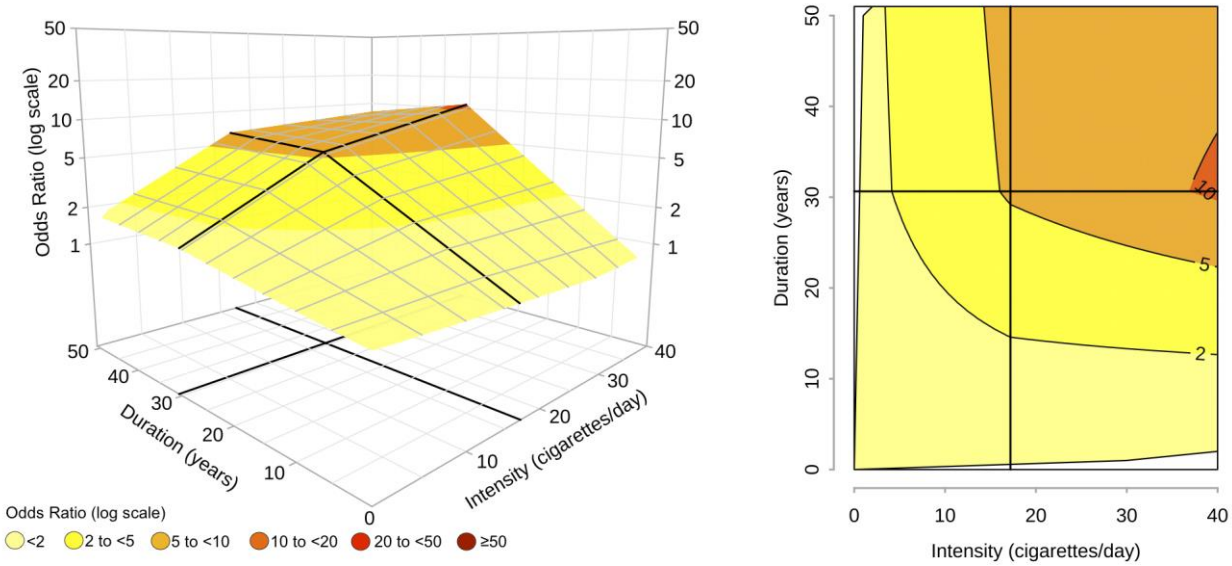
eTable 2 - Odds ratios (ORs)[a] and 95% credible intervals (CIs) of head and neck cancer in former smokers who quit ≥10 years ago, by cancer type, duration (years) and intensity (cigarettes/day) of cigarette smoking estimated through step-function as compared with results from bivariate spline models. INHANCE consortium

	Intensity (cigarettes/day)					
	1-15		16-25		26-40	
Cancer type	**Ca:Co**	**OR (95%CI) [Min-Max][b]**	**Ca:Co**	**OR (95%CI) [Min-Max][b]**	**Ca:Co**	**OR (95%CI) [Min-Max][b]**
Oral cavity and pharynx						
Duration (years)						
1-25	404:2268	1.0 (0.9-1.0) [1.0-1.4]	293:1323	1.2 (1.1-1.4) [0.9-1.6]	174:656	1.4 (1.3-1.6) [0.9-1.9]
26-35	111:435	1.4 (1.2-1.7) [1.3-1.7]	152:515	1.7 (1.5-1.9) [1.5-2.0]	105:281	1.9 (1.5-2.3) [1.7-2.5]
36-51	83:224	1.7 (1.4-2.2) [1.4-2.2]	111:232	2.3 (1.9-2.8) [1.8-2.8]	89:154	2.5 (2.0-3.2) [2.1-4.0]
Larynx						
Duration (years)						
1-25	107:2164	1.4 (1.3-1.5) [1.0-2.9]	113:1256	2.7 (2.4-2.9) [1.2-4.2]	51:627	2.3 (1.9-2.7) [1.0-4.4]
26-35	66:418	3.2 (2.7-3.8) [1.9-4.3]	109:494	4.6 (4.0-5.1) [3.2-6.7]	54:264	4.1 (3.3-5.0) [3.0-7.4]
36-51	50:224	3.2 (2.5-3.9) [2.3-7.9]	91:223	6.3 (5.3-7.3) [4.7-14.7]	70:134	7.7 (6.2-9.5) [4.5-16.5]

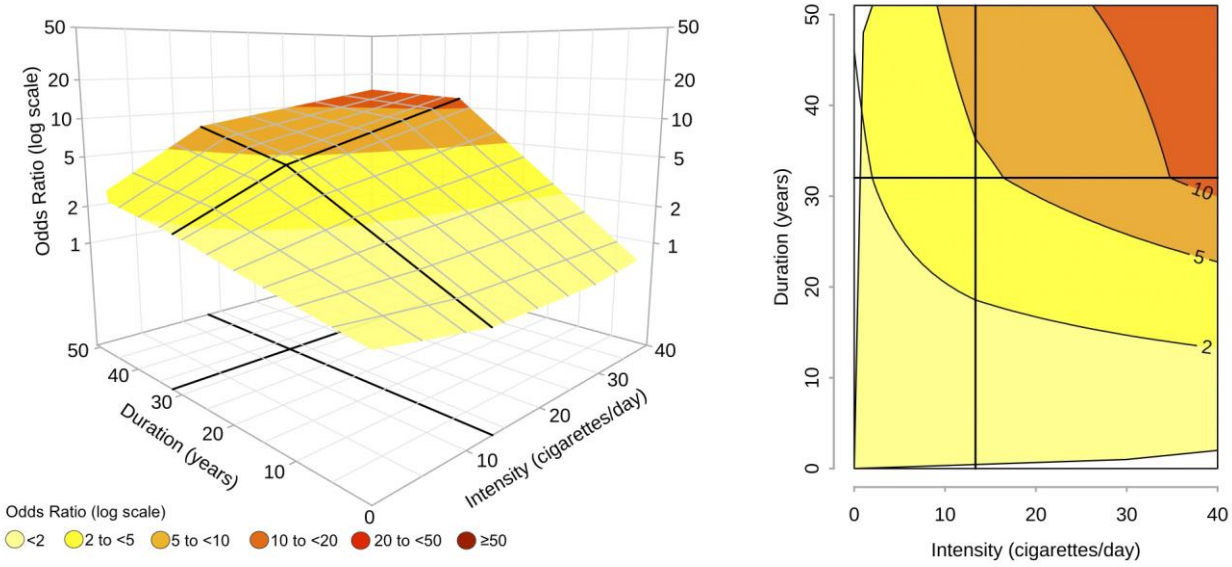
[a] Fitted models included adjustment for age, sex, race, study, education, drinking status, drinking intensity, and drinking duration. The reference category was defined as ‘‘Never smokers’’. The reference category included 2791 cases and 13,139 controls for the analysis on oral and pharyngeal cancer and 330 cases and 11,403 controls for that on laryngeal cancer. [b] Min and Max represent the lowest and the highest OR values estimated for any combinations of intensity and duration by bivariate spline models.

eFigure 2 - Odds ratios[a,b] of oral cavity and oropharyngeal cancers in <u>current smokers,</u> for the joint effect of intensity (cigarettes/day) and duration (years) of cigarette smoking estimated through bivariate spline models. INHANCE consortium

A. Oral cavity cancer



B. Oropharyngeal cancer

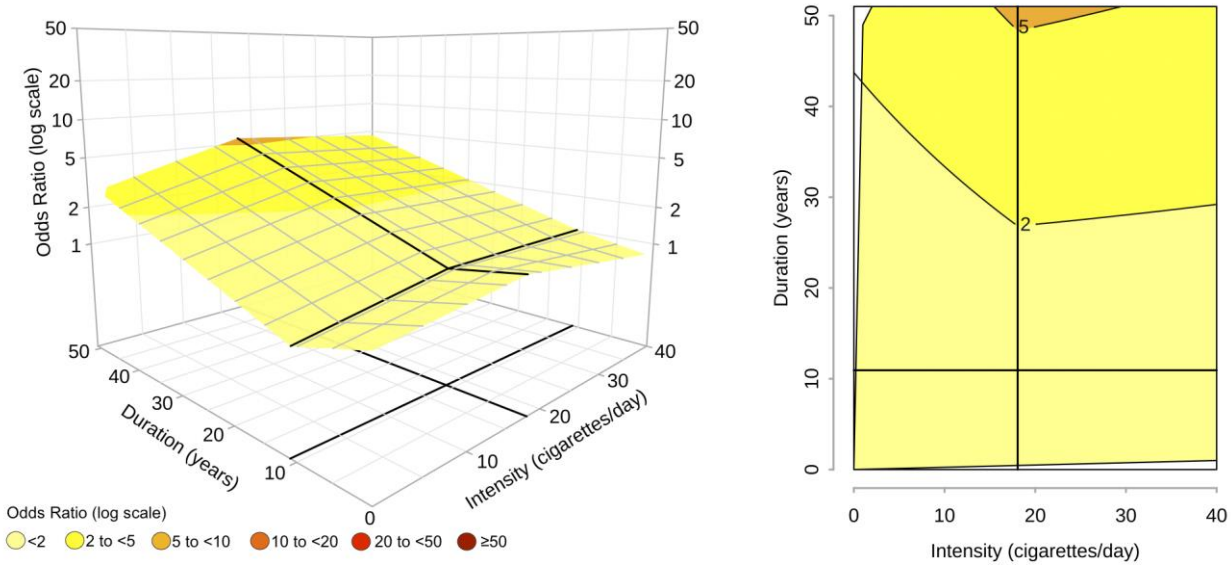


[a] Fitted models included adjustment for age, sex, race, study, education, drinking status, drinking intensity, and drinking duration. The reference category was defined as "Never smokers".

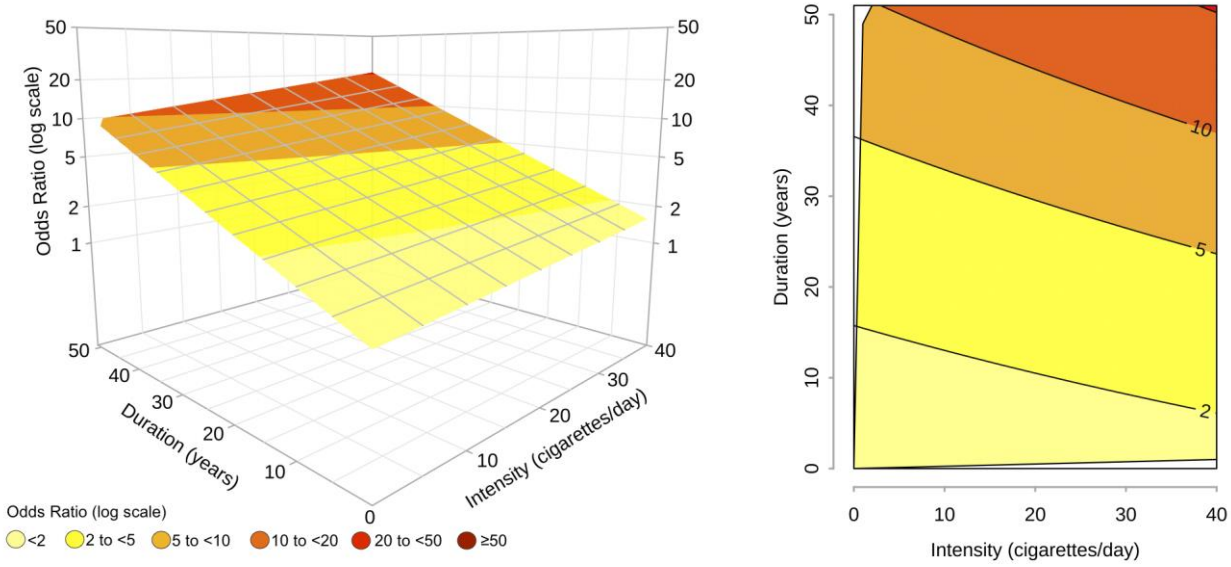
[b] On the grid, black thicker lines represent knot locations: 17 cigarettes/day and 30 years of duration for oral cavity cancer and 14 cigarettes/day and 32 years of duration for oropharyngeal cancer, respectively. Dark grey lines in contour plots (right) indicate iso-risk curves at defined risk levels.

eFigure 3 - Odds ratio[a, b] of oral and pharyngeal cancer and laryngeal cancer in <u>former smokers (no matter the time they quit)</u>, for the joint effect of intensity (cigarettes/day) and duration (years) of cigarette smoking estimated through bivariate spline models. INHANCE consortium

A. Oral and pharyngeal cancer



B. Laryngeal cancer



[a] Fitted models included adjustment for age, sex, race, study, education, drinking status, drinking intensity, and drinking duration. The reference category was defined as "Never smokers".

[b] On the grid, black thicker lines represent knot locations: 18 cigarettes/day and 11 years of duration for oral cavity and pharyngeal cancer. Dark grey lines in contour plots (right) indicate iso-risk curves at defined risk levels.