Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

DISEÑOS EN EPIDEMIOLOGÍA

MÓDULO II

Tema 1

Diseño y atributos de los estudios de casos y controles

Arem H, Neuhouser ML, Irwin ML, et al. Omega-3 and omega-6 fatty acid intakes and endometrial cancer risk in a population-based case-control study. *Eur J Nutr*. 2013;52(3):1251-1260. doi:10.1007/s00394-012-0436-z

Materials and methods

Study design

A population-based case–control study was conducted in Connecticut, involving English-speaking residents aged 35–81 years who were diagnosed with incident, primary endometrial cancer between October 2004 and September 2008. Study design and eligibility have been described elsewhere [5]. Of the 1,216 potentially eligible patients identified statewide through the Rapid Case Ascertainment Shared Resource of the Yale Cancer Center, 317 chose not to participate, 19 had died before study contact, 13 were too ill, 44 could not be located, 68 could not be reached by telephone and 87 were ineligible. Among 1,995 Connecticut women in the eligible age range identified as potential controls through random digit dialing, 1,447 agreed to further contact for participation and 1,248 were contacted, while 111 were ineligible due to residence, mental impairment, language barrier, cancer diagnosis or ineligible medical conditions. Another 92 women were disqualified due to illness or residence outside of Connecticut, and 371 refused to participate. Research staff enrolled 668 (54.9 %) of diagnosed endometrial cancer cases and 674 (64.5 %) of contacted, eligible controls. In person interviews were carried out at participant homes. After completion of signed informed consent, study staff administered structured questionnaires on ethnic and demographic factors, environmental exposures and lifestyle factors. The study was approved by the Institutional Review Boards of Yale University, the Connecticut Department of Public Health Human Investigation Committee and the 28 participating Connecticut hospitals.

Exposure assessment

Diet was assessed using a self-administered 120-item food frequency questionnaire (FFQ) from the Fred Hutchinson Cancer Research Center. This FFQ was modified from the Women's Health Initiative FFQ, and was validated against 4-day food records and 24-h dietary recalls with published measurement characteristics [28]. Participants completed the mailed questionnaire, which was reviewed by research staff during the home visit. The FFQ inquired about the frequency and portion size of various foods based on estimated usual intake over the previous 1–5 years and >19 adjustment questions queried about types and quantities of fat used in cooking and at the table. These responses were applied to analysis algorithms to normalize calculated fat intakes. Participants were specifically asked about consumption of dark fish (such as salmon, mackerel or blue-fish), white fish (such as sole, halibut, snapper or cod), shellfish (such as shrimp, lobster, crab or oysters) and fried fish on separate line items. Also, they were asked whether they took fish oil, omega-3 or cod liver oil in the 1–5 years prior and if so, they were asked to choose a

category for frequency of supplementation (<1/week, 1–2 days/week, 3–4 days/week, 5–6 days/week, 7 days/week). The primary nutrient density database for this FFQ was derived from the Nutrition Data Systems for Research (NDS-R, version 2008, Nutrition Coordinating Center, University of Minnesota, Minneapolis, MN) and has been augmented with information from manufacturers. Cases were asked to recall average diet in the period 1–5 years prior to diagnosis so as to minimize dietary changes occurring because of disease; controls were asked to recall average diet 1–5 years prior to interview.

Statistical analysis

Control women who had a hysterectomy (n = 6) or were outside the specified age range (n = 3) were excluded from analyses. We excluded study subjects who missed >10 items on the FFQ (n = 118 cases and 89 controls) as these FFQs were considered unreliable. We also removed subjects whose calculated energy intake was less than 600 (n = 28) or above 5,000 kcal per day (n = 9) on the FFQs as these cutoffs have been used in similar populations to eliminate subjects whose FFQ information is believed to be inaccurate or unrepresentative of usual diet [29, 30]. Our final analytic sample size was 556 cases and 533 controls. We performed descriptive analyses using the t-distribution for continuous variables and Chi-squared distribution for categorical variables. Total n-6 fatty acids included linoleic acid (18:2) and arachidonic acid (20:4). n-3 fatty acids were summed as the total of linolenic acid (18:3), eicosapentaenoic acid (EPA, 20:5) docosahexaenoic acid (DHA, 22:6), and docosapentaenoic acid (22:5). Individual fatty acids were divided into quartiles based on intake among controls. Total fish consumption was calculated as the sum of reported fried, dark, white and shellfish weekly servings. Because fish consumption in our study population was low for specific types of fish, we analyzed fried, dark, white and shellfish individually into categories of ever (>0 servings/year) or never (0 servings/year).

We examined the correlation between individual and grouped fatty acids and report selected Spearman correlation coefficients as follows: total n-6 PUFAs and linoleic acid = 0.99, DHA and EPA = 0.95, docosapentaenoic acid and DHA = 0.91, docosapentaenoic acid and EPA = 0.95. Given the high correlations between docosapentaenoic acid and the long-chain fatty acids DHA and EPA, very low absolute intake of docosapentaenoic acid and an insufficient body of evidence supporting a role for docosapentaenoic acid in carcinogenesis, docosapentaenoic acid was not analyzed for its main effect, but rather included in total n-3 fatty acids.

To account for differences in fatty acid intake due to differences in total energy we used the multivariate nutrient density method, dividing fatty acid intake by total energy and multiplying by 1,000 [31]. After assessing differences in quartiles using the log rank test, to estimate odds ratios (ORs) and 95 % confidence intervals (CIs) we built logistic regression models. All variables in Table 1 were examined for possible confounding. We retained variables significant at the two-sided p = 0.05 level, those that caused a>10 % change in odds ratio estimates and variables that were selected a priori to be included in the model based on previously observed associations with risk. For final statistical models, we included adjustment for age at interview (continuous), race (white

or other), body mass index (continuous), number of live births (continuous), menopausal status (yes/no), oral contraceptive use (ever/never), smoking category (never, former, current), and physician-diagnosed hypertension (yes/no). Education, diabetes, vegetable consumption and physical activity levels were considered but were not included in the multivariate-adjusted models because adding these variables to the models did not change parameter estimates by >10 %. We performed linear trend tests by assigning values of 1–4 for each quartile and treating the ordinal variable as continuous. To maximize power, we also created continuous models scaling the exposure of interest by the interquartile range.

Aschengrau A, Gallagher LG, Winter M, Butler L, Patricia Fabian M, Vieira VM. Modeled exposure to tetrachloroethylene-contaminated drinking water and the occurrence of birth defects: a case-control study from Massachusetts and Rhode Island. Environ Health. 2018 Nov 6;17(1):75. doi: 10.1186/s12940-018-0419-5. PMID: 30400949; PMCID: PMC6219161.

Methods

Selection of study population

Cases were comprised of live- and stillborn infants who were delivered from 1968 through 1995 to residents of 24 MA and four RI cities and towns with some VLAC water distribution pipes. Approximately 480 miles of VLAC pipes were installed in these towns, representing approximately 63% of the VLAC pipes in the two states. The remaining RI and MA towns with VLAC pipes were excluded from the present study because they had few VLAC pipes, lacked documentation on the locations and dates of VLAC pipe installation, and/or small resident populations. Available water sampling data indicated that PCE contamination persisted in public water supplies of selected towns through the 1990s because the target level for remediation was 40 µg/L [7].

Cases were identified by abstracting birth defect diagnoses from livebirth, fetal death, and death certificate records (*N* = 479). Based on associations observed in our prior cohort study [20], infants with the following defects were selected: central nervous system defects (CNS) (including anencephaly, spina bifida with or without hydrocephalus, encephalocele, hydrocephalus alone, microcephaly and other CNS defects), oral clefts (cleft lip with and without cleft palate and cleft palate alone), and hypospadias. Cases with more than one birth defect were eligible for inclusion; however, those with recognized syndromes were excluded. Prenatal testing followed by elective termination of affected pregnancies was uncommon during the case ascertainment period [29, 30].

Controls were randomly selected from livebirth records of non-malformed infants born during the same period to residents of the same geographic area as the cases. The control selection process was stratified by state and birth year so that the number of controls selected from MA and RI was

proportional to their number of births over the ascertainment period (45% from RI and 55% from MA). A total of 800 controls were targeted for selection; 794 remained after excluding duplicate subjects.

Questionnaire and vital record data collection

Paper copies of livebirth, fetal death and death certificates and computerized vital record data were abstracted to obtain parent's and infant's names; maternal address at delivery; infant's date of birth; maternal age, race, and educational level; paternal age and educational level; maternal pregnancy history; date of last menstrual period; prenatal care information and birth defect diagnoses.

Mothers were subsequently traced and sent self-administered questionnaires. Overall, 4.2% of case mothers (*N* = 20) and 7.2% of control mothers (*N* = 57) were found to be deceased. We successfully located 87.4 and 88.3% of living case (*N* = 401) and control mothers (*N* = 651), respectively, and, of these, 38.6% of case mothers (*N* = 155) and 31.8% of control mothers (*N* = 207) returned the questionnaire after two mail and one telephone reminder. The purpose of the questionnaire was to identify mothers who moved during pregnancy and augment vital records data on birth defect diagnoses and confounding variables.

When we compared the demographic characteristics of questionnaire respondents and non-respondents, we found that the two groups were similar with respect to PCE exposure status (68.5% of case respondents versus 68.0% of case non-respondents and 66.2% of control respondents versus 70.7% of control non-respondents) and delivery year (e.g., 28.9% of case respondents versus 26.7% of case non-respondents and 37.4% of control respondents versus 37.7% of control non-respondents delivered during 1968–1978). While respondents were more likely to reside in Massachusetts (77.2% of case respondents versus 66.8% of case non-respondents and 72.8% of control respondents versus 51.0% of control non-respondents), be older at delivery (median age was 29.0 for case respondents versus 26.7 for case non-respondents and 28.3 for control respondents versus 26.6 for control non-respondents), and begin prenatal care in the first trimester (93.9% of case respondents versus 85.3% of case non-respondents and 90.8% of control respondents versus 86.3% of control non-respondents), these differences were present for both case and control mothers who returned our study questionnaire.

Geocoding of residential addresses

Residential addresses at birth recorded in vital records were geocoded to a latitude and longitude using ArcGIS (10.0, ESRI, Redlands, CA). First trimester addresses of questionnaire respondents who moved during pregnancy were also geocoded (*N* = 33 cases and 33 controls). Whenever possible, each address was assigned to a parcel of land. Addresses that could not be geocoded to a parcel were geocoded to the closest parcel address by street number. If a street number was unavailable, the address was geocoded to the middle of the street when the street was less than a

mile long or to the intersection of the address with the cross-street when the street was a mile or longer. Overall 98.7% of the addresses were successfully geocoded. All geocoding was conducted without knowledge of case-control status.

PCE exposure assessment

Because historical water sampling data for PCE was scant, we estimated each subject's prenatal exposure used a leaching and transport model. The model was developed for our research by Webler and Brown [31, 32] to estimate the mass of PCE entering the drinking water using the starting quantity of PCE in the pipe liner, the pipe's age, the leaching rate of PCE from the liner into the water, and the water flow rate through the pipe. The initial amount of PCE in the liner was based on the pipe dimensions and information from the manufacturer on the application of the liner. The leaching rate of PCE was determined in experiments by Demond [5].

We incorporated the Webler Brown algorithm into the open source code of EPANET, water distribution system modeling software created by the US EPA [33]. Initially designed to investigate water quality problems, EPANET has been used in several epidemiological studies investigating adverse health effects of drinking water contaminants (e.g. [34]). The combined modeling approach integrated pipe schematics, water use, and PCE transport to determine the water flow rate and direction and the amount of PCE at points of consumption throughout the distribution system.

GIS maps of geocoded residences and the water distribution systems were used to construct a graphic for each town depicting the locations and characteristics of the pipes (e.g. VLAC or not) and consumption points. We assigned each mother's residence to the closest consumption point on the distribution system. The model simulation estimated the average mass of PCE in grams delivered to each subject's residence during the calendar year when the first trimester ended. We estimated the annual exposure because we did not have data on the month of move-in or VLAC pipe installation. Average monthly exposure was obtained by dividing annual exposures by 12. Questionnaire respondents who report using a private well for their prenatal water supply (N = 21) were considered unexposed.

Data analysis

The strength of the association between PCE exposure and each birth defect was estimated with odds ratios (OR) and statistical stability was evaluated with 95% confidence intervals. The following defect groups were examined: any central nervous system defect, any neural tube defect, any anencephaly, any spina bifida (with and without hydrocephalus), any other CNS defects, any oral cleft, any cleft lip (with or without cleft palate), cleft palate alone, and hypospadias. Cases with more than one of these defects contributed to each subgroup. Analyses of hypospadias were limited to male cases and controls. ORs were calculated only if there were at least three exposed cases and three exposed controls.

Analyses first compared mothers who were ever exposed to PCE-contaminated drinking water during the calendar year the first trimester ended to unexposed mothers. We focused on exposure during the first trimester because the structural defects under study form during this period [35]. Next, we dichotomized PCE exposure at 1.136 g, the level corresponding to an average monthly drinking water concentration of 40 µg/L during the calendar year that the first trimester ended. This concentration was the criterion for remediation when PCE contamination was discovered in 1980. Levels above 40 µg/L were designated as "high." Each set of analyses was repeated after (1) incorporating exposure levels at the first trimester address for questionnaire respondents who moved during pregnancy, (2) excluding cases and controls with a family history of defects, (3) excluding cases with multiple defects, and (4) excluding cases included in our prior cohort study (*N* = 10).

Logistic regression models estimated ORs while controlling for confounding variables. We selected potential confounders from those available in the vital records and self-administered questionnaires based on a literature review and construction of directed acyclic graphs. Multiple imputation was used to obtain values of potential confounders with missing data. The amount of missing data ranged from 0% (state of residence and delivery year) to 72.5% (maternal alcoholic beverage consumption). Variables with a high proportion of missing data came solely from the self-administered questionnaires. Twenty imputed data sets were generated using the fully conditional specification (FCS) multiple imputation method based on 27 variables. Point and variance estimates from the imputed data sets were subsequently combined and used in adjusted analyses.

Initial adjusted models controlled for one potential confounder at a time. Controlling for delivery year had the greatest impact on the crude associations and so it was included in all multivariate models. We also decided to include state in all multivariable models to account for variations in case ascertainment between the two states. Additional variables that changed the crude association between PCE exposure and each type of birth defect by > = 10% were also included in the final models; these included maternal alcoholic beverage consumption in the oral cleft analyses and maternal race in the hypospadias analyses.

D'Souza G, Kreimer AR, Viscidi R, et al. Case-control study of human papillomavirus and oropharyngeal cancer. *N Engl J Med*. 2007;356(19):1944-1956. doi:10.1056/NEJMoa065497
Methods

Patients

Our case–control study was nested within a longitudinal cohort study of patients with newly diagnosed squamous-cell carcinomas of the head and neck in the outpatient otolaryngology clinic of the Johns Hopkins Hospital in Baltimore from 2000 through 2005. Eligible case patients included those with a confirmed diagnosis of oropharyngeal squamous-cell carcinoma.

The control group consisted of patients without a history of cancer who were seen for benign conditions between 2000 and 2005 in the same clinic from which the case patients were enrolled (Table 1). Subsequent to enrollment of a case, eligible control patients within the same sex and 5-year age categories were approached until two control patients were individually matched to each case patient. The study protocol was approved by the institutional review board of the Johns Hopkins Hospital. Written, informed consent was obtained from all patients.

Data Collection

Specimens were collected from case patients before therapy and from control patients at enrollment. Oral-mucosal specimens were collected with the use of a saline oral rinse and 5 to 10 strokes of a cytology brush (Oral CDx, CDx Laboratories) on the posterior oropharyngeal wall. Serum samples were collected and stored at −80°C. For case patients, formalin-fixed, paraffin-embedded tumor specimens and, if possible, snap-frozen fresh tumor specimens were obtained for the detection of HPV.

All patients completed an audio, computer-assisted self-administered interview that obtained information about demographic characteristics, oral hygiene, medical history, family history of cancer, lifetime sexual behaviors, and lifetime history of marijuana, tobacco, and alcohol use (see the Supplementary Appendix, available with the full text of this article at www.nejm.org).

Laboratory Studies

In Situ Hybridization for HPV-16 Detection

We looked for HPV-16 in formalin-fixed and paraffin-embedded tumors from all case subjects, using in situ hybridization–catalyzed signal amplification for biotinylated probes (Dako GenPoint).[15] The HPV-16-positive status of a tumor was defined as specific staining of tumor-cell nuclei for HPV-16.

DNA Purification and Analysis

DNA from oral specimens[16] and fresh-frozen tumors[17] from a subgroup of case subjects was purified as previously described. The tumor specimens were microdissected to ensure that more than 70% of the sample was DNA from the tumor.

We analyzed purified DNA for 37 types of HPV by means of a multiplex polymerase-chain-reaction (PCR) assay targeted to the L1 region of the viral genome, using PGMY09/11 L1 primer pools and primers for β-globin, followed by hybridization to a linear probe array (Roche Molecular Systems).[18] The HPV-16 viral load in purified DNA from oral-mucosal specimens and fresh-frozen tumor specimens was determined with the use of a sensitive real-time PCR assay targeted to the E6 coding region.[16,19] The viral load was reported for positive samples (those with ≥1 copy of the

virus) and was adjusted to the total number of human cells tested with the use of a real-time PCR assay targeted to a single copy of a human gene (for endogenous retrovirus 3, *ERV3*).[16]

Serologic Analysis

Serum antibodies to the HPV-16 L1 protein were detected with the use of an enzyme-linked immunosorbent assay (ELISA) based on virus-like particles.[20] Antibodies against HPV-16 E6 and E7 oncoproteins were detected with the use of ELISA and bacterially expressed full-length E6 or E7 as the antigen.[21]

Statistical Analysis

Cumulative alcohol use was calculated as follows. We defined a drink-equivalent as one 12-oz beer, one 6-oz glass of wine, one 3-oz mixed drink, or one 1.5-oz shot of liquor. The number of drink-equivalents per week was determined for each patient within each 5-year age interval and combined into a measure of lifetime alcohol use, defined as the number of years during which 15 or more drink-equivalents (hereafter called "drinks") per week were consumed.

We calculated cumulative tobacco use in pack-years using information about the frequency of use (number of cigarettes, pipes, or cigars smoked per day) and duration of use (during 5-year age intervals) and accounting for gaps in use. Four cigars or five pipes per day were deemed equivalent to one pack of cigarettes in the calculation of pack-years.[22]

Unconditional and conditional multivariate logistic-regression models were used to estimate odds ratios and the associated 95% confidence intervals (CIs). Results from the unconditional and conditional models were similar, and the results from the unconditional models are presented. Final multivariate models were created through stepwise elimination of variables of interest from univariate analysis while biologically relevant variables were retained. Owing to the colinearity of sexual behaviors, the effect of each behavior on the risk of cancer was evaluated in separate multivariate models adjusted for alcohol use, tobacco use, presence or absence of a family history of head and neck cancer, oral hygiene, age, and sex. To evaluate trends in odds, ordinal variables were modeled as single, continuous, independent variables. Multiplicative interactions among exposure to HPV, tobacco use, and alcohol use were evaluated by including an interaction term in the regression model, and statistical significance was determined with the use of the likelihood-ratio test. For comparison of our results with those in previous reports,[9,10] additive interactions were evaluated with the use of a synergy index, calculated as (odds ratio for tobacco or alcohol use and HPV−1)÷([odds ratio for tobacco or alcohol use+odds ratio for HPV]−2).[23] The odds ratio for HPV was for either seropositivity or infection. Attributable risk was calculated as previously described.[24] P values of less than 0.05 for associations were considered to indicate statistical significance. Stata 8.0 software (Stata) was used for all analyses.

Mork J, Lie AK, Glattre E, et al. Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. *N Engl J Med*. 2001;344(15):1125-1131. doi:10.1056/NEJM200104123441503

Methods

Subjects and Study Design

Almost 900,000 residents of Norway, Finland, and Sweden have donated serum samples to the four serum banks participating in the study (additional information is in the Appendix [available only with the electronic version of the article]).

Persons who had donated serum at least one month before a diagnosis of a head or neck cancer were identified by linkage of serum-bank files with the national cancer registries in Norway, Finland, and Sweden. Reporting of new cases of cancer is compulsory in these three countries, and reliance on multiple data sources ensures that the cancer registries are almost 100 percent complete.[14]

Head and neck sites were defined according to the following codes of the *International Classification of Diseases, Seventh Revision* [15]: 140 (vermilion border of the lips), 141 (tongue), 143 (floor of mouth), 144 (oral cavity, not otherwise specified), 145 (oropharynx), 146 (nasopharynx), 147 (hypopharynx), 148 (pharynx, not otherwise specified), 160 (nose and paranasal sinuses), and 161 (larynx).

From the creation of the serum banks through 1997, 301 invasive squamous-cell carcinomas and 8 carcinomas of the head and neck (not otherwise specified) were registered. Reevaluation of pathological and clinical features led to the exclusion of four cases because the histologic diagnosis was uncertain and two cases because their true anatomical location was outside the sites designated for the study. Of the eight cases of carcinoma not otherwise specified, two cases reclassified as squamous-cell carcinoma were included, and the other six unspecified carcinomas were excluded. In five cases, serum samples were not available. The characteristics of the remaining 292 patients are given in Table 1. If more than one prediagnostic serum sample was available, the first (oldest) sample was chosen. The mean time between enrollment and diagnosis was 9.4 years (range, 2 months to 19.3 years).

Characteristics of the Patients with Head and Neck Cancer, According to Cohort.

For each patient, five (Norway and Sweden) or seven (Finland) matched control subjects were selected. The controls were alive and free of head and neck cancer at the time the corresponding patient received a diagnosis of cancer. The matching variables were sex, age at the diagnosis of cancer in the corresponding patient (within two years), and length of serum storage (within two months). Matching of patients and controls was performed entirely within each cohort (serum bank) to ensure that differences between the cohorts did not affect the validity of the study. In

Norway, the patients and controls were also matched according to county of residence. If five matched control subjects per patient could not be found, the matching criteria for age and serum storage time were expanded stepwise by one year of age and two months of serum storage. The mean difference in age between patients and controls was 0.9 year, and the maximal difference was 4 years. The mean difference in serum storage time was 0.8 month, and the maximal difference was 6 months. After the exclusion of 22 eligible controls for whom serum samples were not available, the control group contained 1568 persons. There were at least four matched controls for each patient. Diagnostic histologic specimens from 228 of 292 patients were received from pathological laboratories for histopathological review and polymerase-chain-reaction (PCR) analysis.

Laboratory Methods

Antibodies against HPV were detected by the standard enzyme-linked immunosorbent assay, with the use of baculovirus-expressed capsids containing both the L1 and the L2 proteins (major oncogenic HPV-16, HPV-18, and HPV-33) or only L1 (HPV-73). HPV-73 has been cloned from an esophageal carcinoma.[16] Disrupted capsids of bovine papillomavirus served as a negative control. The cutoff levels used to assign seropositivity from continuous absorbance values were preassigned and, relative to internal standard serum, were the same as those used in previous studies.[5,12,13,17,18] For the different viruses, the interassay coefficients of variation ranged from 17.8 percent to 33.8 percent, and the intraassay coefficients of variation ranged from 5.3 percent to 10.4 percent.

Serum cotinine, a biochemical marker of exposure to tobacco smoke,[19] was measured by a quantitative competitive enzyme immunoassay that used microtiter plates coated with anticotinine antibodies and detection with a cotinine–horseradish peroxidase conjugate (STC Technologies, Bethlehem, Pa.). On the basis of previous reports,[19–21] prospectively chosen cutoff levels of serum cotinine were used to identify nonsmokers and those "passively exposed to smoke" (0 to 19.99 ng of cotinine per milliliter), "light to moderate smokers" (20.00 to 224.99 ng of cotinine per milliliter), and "heavier smokers" (≥225.00 ng of cotinine per milliliter).

Formalin-fixed, paraffin-embedded tissues were examined for HPV DNA by PCR assay. The quality of DNA was tested by amplification of HLA-DQA1 with the primers GH26 and GH27.[22] All samples from patients with cancer that were positive for these primers were examined for HPV DNA with the L1 consensus primers GP5+ and GP6+[23] and the E1 consensus primers CpI and CpIIG,[24] as previously described.[25] Empty paraffin-block sections were cut between samples from patients with cancer and used as contamination controls for each PCR assay. HPV-DNA–positive samples were tested with E6 and E7 type-specific primers for HPV-6, HPV-11, HPV-16, HPV-18, and HPV-33.[25–27] All samples that were negative for HPV DNA were also tested with primers specific for HPV-16.

All laboratory analyses were performed with masked samples, and then the data were submitted to the Cancer Registry of Norway for decoding and statistical analysis.

Statistical Analysis

Odds ratios and their 95 percent confidence intervals were derived from conditional logistic-regression models with the Epicure program.[28] The logistic-regression analyses reflected the three matching variables of sex, age, and length of serum storage. Likelihood-ratio tests evaluated variables in the model, including a test of homogeneity in odds ratios. Pearson's correlation coefficient estimated the correlation between variables. Fisher's exact test was used to test for equity between proportions. A two-tailed P value of less than 0.05 was considered to indicate statistical significance.

Anantharaman D, Muller DC, Lagiou P, et al. Combined effects of smoking and HPV16 in oropharyngeal cancer. *Int J Epidemiol*. 2016;45(3):752-761. doi:10.1093/ije/dyw069

Materials and Methods

Study sample

This analysis included two studies of HPV serology and HNC, the Alcohol-Related Cancers and Genetic Susceptibility in Europe (ARCAGE) study and the HNC case-control study nested within the European Prospective Investigation Into Cancer and Nutrition (EPIC) cohort. Briefly, the ARCAGE study was conducted during 2002–05 and included 1292 pathologically confirmed primary HNC and 1425 controls frequency-matched for age, sex and area of residence. [11,16] Ever smokers were defined as individuals who smoked any tobacco product at least once a week for 1 year, and ever drinkers were those who reported ever consuming any alcoholic beverage. [17] The EPIC cohort recruited 521 330 individuals during 1992 and 2000, of whom 385 747 participants contributed a blood sample. [18] This analysis included 612 incident HNC and 1599 controls. [12] Two controls (one in Denmark) were randomly selected for each cancer patient from appropriate risk sets consisting of all cohort participants alive and free of cancer (except non-melanoma skin cancer) at the time of diagnosis of the index case. Controls were matched on country, sex, date of blood collection (1 month, relaxed to 5 months for sets without available controls) and date of birth (1 year, relaxed to 5 years for sets without available participants). Ever smokers were individuals who reported ever smoking any tobacco product in their lifetime, and ever drinkers were individuals who reported ever consuming any alcoholic beverage. HNC included cancers arising at the oral cavity (International Classification of Diseases for Oncology (ICD-O) C00.3–C00.9, C02.0–C06.9, C14.0–C14.9, excluding C02.4, C02.8, C02.9, C05.1, C05.2, C05.8, C05.9), oropharynx (ICD-O: C01, C02.4, C05.1–C05.2, C09, C10), hypopharynx and larynx (ICD-O: C13, C32) and non-specified and overlapping sites (ICD-O: C02.8, C02.9, C05.8, C05.9, C32.8). Lymphomas were not included, and salivary gland cancers were omitted. This analysis included head and neck cancers of all histological subtypes, of which squamous cancers comprised the vast majority (~ 91%), and some other rarer non-squamous histologies (6%, in ARCAGE and 9% in

EPIC). Informed consent was obtained from all participants in both the studies, and the studies were approved by the ethical review boards at the participating centres and the International Agency for Research on Cancer.

HPV serology

HPV antibodies were assayed using the bead-based multiplex serology method as described elsewhere. [19] Testing was performed blind to the case-control status of the participants. Mean fluorescence intensity (MFI) values were dichotomized by applying thresholds derived from a cross-sectional study among Korean students of mean plus 5 standard deviations (SD; for HPV16 E6) or the mean plus 3 SD excluding positive outliers (for HPV16 L1), [20] as described previously. [11,12]

Statistical analysis

The overall associations between HPV16 (L1 and E6), smoking, alcohol intake and HNC risk were assessed by calculating odds ratios (ORs)and their corresponding 95% confidence intervals (CIs).These models included age, sex, smoking status (never, former, current), alcohol consumption (never, ever plus ethanol g/day at recruitment) and country as covariates. Since certain combinations of exposures were very rare (e.g. HPV16 E6-positive never smoking control subjects), Bayesian logistic regression models were used to calculate ORs and corresponding 95% credible intervals (CrI). These models use a prior distribution to shrink or penalize the regression coefficients, thus providing more stable estimates than maximum likelihood methods. Following Gelman *et al.* , [21] all regression inputs were centred, and continuous inputs were re-scaled to have a standard deviation of 0.5. All regression coefficients were then modelled with a weakly informative Cauchy prior distribution with mean 0 and scale 2.5, with the exception of the intercept, which was given a weaker Cauchy prior with scale 10. These models were fitted using the bayesglm function in the R package ARM. [21,22] In these analyses, former and current smokers were combined as ever smokers and given the few participants who reported never consumption of any alcoholic beverage; individuals who consumed 7 g or less of ethanol (equivalent of half a drink) per day were considered the reference. Since the results from ARCAGE and EPIC studies were similar, data were pooled in order to obtain more precise estimates. Interactions between smoking, alcohol intake and HPV16 were examined by the inclusion of an interaction term in the penalized regression models. Additive interactions were evaluated by estimating the synergy index (SI). [23] The prevalence of OPC by categories of smoking and HPV16 were calculated based on the ORs from the fitted models and assumed population prevalence of 0.003, based on the cumulative risk for pharyngeal cancer among men and women combined, in more developed regions of the world. [24] All statistical analyses were performed using Stata version 11.2 (StataCorp, College Station, TX, USA) and R version3.1.0. [25]