UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di dottorato in scienze biomediche, cliniche e sperimentali

Dipartimento di scienze cliniche e di comunità

Dottorato di ricerca in statistica biomedica

Ciclo XXVIII – Settore scientifico disciplinare MED/01

TESI DI DOTTORATO DI RICERCA

“Statistical methods to analyze continuous risk variables in individual patient data meta-analyses: application on a study on tobacco smoking and gastric cancer risk in a consortium of case-control studies (the Stomach Pooling (StoP) Project)”

Dottoranda
Delphine Praud

Tutor
Prof. Carlo La Vecchia

Coordinatore del Dottorato
Prof. Adriano Decarli

2015

CONTENTS

LIST OF FIGURES

LISTE OF TABLES

AKNOWLEDGMENT

Ringraziamenti – Remerciements

Volevo ringraziare i membri della commissione Prof Mario Grassi, Prof. Clelia Di Serio and Prof Carlo La Vecchia di essere presenti alla discussione della mia tesi di dottorato.

Volevo ringraziare il Prof. Adriano Decarli di avermi permesso di seguire questi 3 anni di dottorato in Università e anche di partecipare a diversi congressi e meeting, oltre a un periodo all’estero

Volevo ringraziare il mio relatore, il Prof. Carlo La Vecchia, di avermi proposto di venire a Milano per il dottorato e di avermi coinvolta nel progetto StoP. Grazie anche per la tua gentilezza, la tua accoglienza, il tuo sostegno e il tuo modo di lavorare e di insegnare il lavoro di epidemiologa. Ho imperato tantissimo grazie a te e te ne saro grata per sempre.

Volevo ringraziare il gruppo dello StoP project per avermi inclusa nel progetto.

Volevo ringraziare tutto il dipartimento di Epidemiologia dell’istituto Mario Negri, che ha fatto che ho deciso di venire senza nessun dubbi in Italia, sicura di trovare la disponibilità di Cristina, i consigli di Liliane, la simpatia di Silvano, l’aiuto amministrativo di Ivana, i racconti di Alessandra, la professionalità di Eva, la tranquillità di Claudio, il sorriso di Carlotta e sopratutto l’amicizia di Alessandra, Greta, Eleonora, Federica, Francesca, Matteo F, Matteo M, Matteo R, Marta, Paola, Tiziana, Valentina e Valentina.

Grazie di cuore di avermi insegnato l’italiano e inclusa nella vostra vita in questi tre anni ma anche per i prossimi.... Grazie per il vostro sostegno, la vostra positività, il vostro aiuto nel lavoro ma anche nella vita milanese, grazie per gli aperitivi e le pizze, grazie per la pausa caffè di ogni mattino ma anche le pause tè per te!! Sono felicissima

e fortunata di avervi conosciuto, siete delle persone e un gruppo fantastico, mi mancherete tanto!

Ringrazio anche le mie compagne del dottorato Alessandra, Giò, Elena, Teresa e Tiziana, per il vostro aiuto, i vostri consigli, il gruppo Whatsapp delle Decarline, le serate passate insieme, le risate e le settimane di convegni. Che bello di esserci trovate lo stesso anno, mi mancherete tanto però un weekend Decarline si farà sicuramente anche dall'altra parte delle Alpi!!

Je souhaite ensuite remercier ma famille, mon papa, ma maman, mes sœurs Gaëlle et Marion, mon frère Rémi et mon petit neveu Théophile. Pour votre soutien parfois silencieux mais toujours présent. Merci d'avoir appris Skype et Whatsapp pour l'occasion. Merci d'avoir toujours été fiers de moi et d'avoir toujours estimé le travail que j'ai fait ici. Merci aussi pour votre implication dans cette nouvelle vie, d'être venus me rendre visite, même de l'autre côté de l'Atlantique et d'avoir eu cette passion pour l'Italie en même temps que moi (et l'italien).

Merci à mes amis d'avoir toujours été là malgré la distance, de m'avoir rendue visite, conseillée et soutenue jusqu'au bout !

Un grazie enorme alle mie coinquiline per vostro sostegno, il vostro ascolto, la vostra presenza, le tisane, le risate, le coreografie, i momenti smalti e di essere state le mie migliore amiche italiane!!

Et pour finir un merci infini, à mon Victor d'avoir accepté que je parte, de ne m'avoir jamais fait peser ce choix, d'avoir tout fait pour qu'on ne s'éloigne jamais, d'avoir finalement apprécié l'Italie, d'être venu me voir aussi souvent, de m'avoir suivi à New York et de m'avoir soutenu par ton amour, ta fierté et tes encouragements dans les moments de stress et de découragement.

ABSTRACT

Gastric cancer represents the fifth most common cancer and the third leading cause of cancer death over both sexes worldwide, with almost 1 million cases and over 700 000 deaths estimated in 2012. The presence of Helicobacter Pylori is a key determinant of gastric cancer. However, other factors, including familial, genetic, environmental and social characteristics appear to also have a role in the etiology of this disease. Tobacco smoking has been associated with increased risk of morbidity and mortality from many diseases and for gastric cancer. Various epidemiologic consortia have been established on several cancers but not yet on gastric cancer. A pooled-analysis of worldwide case-control studies may allow to investigate indebt gastric cancer etiology. Particularly, this large dataset will allow us to better investigate life style characteristics including tobacco smoking, in relation to gastric cancer. The Stomach cancer Pooling (StoP) Project is an international epidemiological consortium. The inclusion criteria for study participation are: a case-control study design (including nested case-control analyses derived from cohort study) and an inclusion of at least 80 cases of gastric cancer (including both cardia and non-cardia location). The aim of my project is to conduct a pooled analysis on data from already available international studies, on the role of tobacco smoking in the etiology of gastric cancer in particular, the number of cigarettes per day and the duration of smoking, using adequate statistical approaches.

During the first year of the PhD program, my project was focused on the two-stage analysis. This method is used to analyze meta-analysis and could be applicable in a case of pooled case-control analysis. The first step of the method consists in calculate adjusted study-specific odds ratios (OR) in order to overcome differences across studies in terms of design or population. The second step consists in summarize these study-specific risks using meta-analytic methods which take into account the heterogeneity across studies. During my second year of PhD program, I studied various statistical methods regarding the analysis of non-linear continuous variables. In addition to transform continuous variables in category, I considered more flexible approaches including fractional polynomials. During my third year of PhD program, I focused my research on a way to adapt these latest methods to the analysis of pooled case-control studies. In particular I chose to use factional polynomials in a two-stage

method due to their simple interpretation and also because their estimates can be easily pooled through a two-stage analysis.

The first step analysis is to perform a fractional polynomial for each study. For each value of the power term (or couple of power terms for the second-order fractional polynomials), the second stage of the model is performed. The pooled dose-response relationship is estimated according to a bivariate random-effects model. The estimate of the trend components could be obtained using restricted maximum likelihood (REML) or maximum likelihood (ML) estimation. The second-stage model is fitted to the data considering each combination of the power terms. The best model, denoted by the optimal power combination is defined as the one minimizing the deviance or the Akaike Information Criterion (AIC), a penalized likelihood which takes into account the number of parameter.

We analyzed data on 21 studies including 10,040 cases and 25,602 controls. To investigate the relationship between tobacco smoking and gastric cancer risk, we first used a classical method, building categories of smokers 1) in terms of quantity; "never smokers", "<10 cigarettes per day", "Between 11 to 20 cigarettes per day", ">20 cigarettes per day" and 2) in terms of smoking duration; "never smokers", "<10 year of smoking", "Between 11 and 30 years of smoking", ">30 years of smoking". We analyzed these variable with a two-stage method. This risk significantly increase with the number of cigarettes per day to reach an OR of 1.29 (95% CI 1.06-1.57)for smokers of more than 20 cigarettes and, with duration to reach an OR of 1.32 (95% CI 1.17-1.49) for smokers smoking for more than 30 years compared to never smokers. These effects of increasing risk are confirmed by different statistical models of analysis including linear model and fractional polynomials, considering the number of cigarettes per day and the duration as a continuous variable.

Results from our analysis confirm that there is an association between cigarette smoking and gastric cancer risk. This risk increases with the number of cigarettes and the duration of smoking. These effects of increasing risk are confirmed by different statistical models of analysis including linear models and fractional polynomials, considering the number of cigarettes per day and the duration as continuous variables.

To our knowledge this is the first study using fractional polynomials through a two-stage random effect methods for pooled case-control studies. Through this method we were able to take into account study-specific adjustment variables and heterogeneity across studies thanks to mixed effect modeling. Categorization has the advantage of a simple epidemiologic interpretation and presentation result. However it assumes that the relationship between the risk of disease and the exposure is flat within intervals and also that there is a discontinuity in response when a category cutpoint is crossed, which is unlikely realistic. Considering exposure variables may avoid these limitations. The relationship between cigarette smoking and gastric cancer risk may be discerned from the categorical analysis, but the analysis of the variable in continuous through polynomials brought additional information in particular to understand the possible threshold and possible changes in slopes.

I. INTRODUCTION AND BACKGROUND

Gastric cancer represents the fifth most common cancer and the third leading cause of cancer death over both sexes worldwide, with almost 1 million cases and over 700,000 deaths estimated in 2012[1]. Gastric cancer incidence rates vary widely across different regions of the world and between men and women. The majority of gastric cancers were reported in developing countries (about 700,000 cases and 550,000 deaths). The highest age-standardized incidence rates (ASR) for gastric cancer were found in Eastern Asia (24.2 per 100 000), Central and Eastern Europe (13.5 per 100 000) and South America (10.3 per 100 000) and the lowest in North-America and in Africa. Comparing genders, rates are 2- to 3-folds higher in men than women (worldwide ASR incidence respectively 17.4 and 7.5 per 100 000 in 2012) [1].

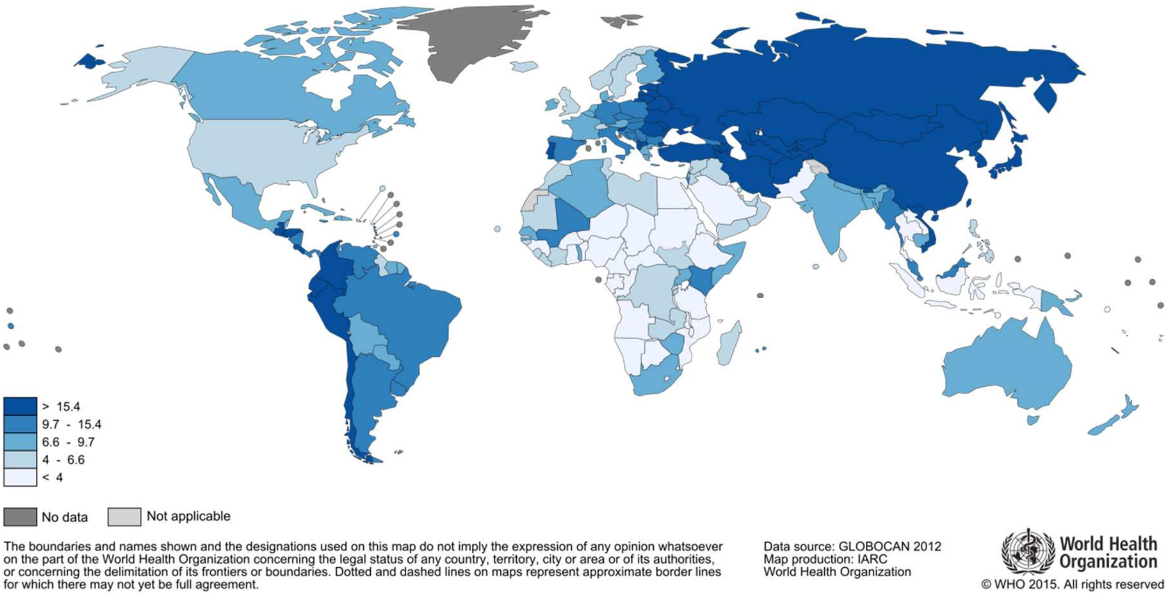

> 15.4
9.7 - 15.4
6.6 - 9.7
4 - 6.6
< 4
No data
Not applicable
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.
Data source: GLOBOCAN 2012
Map production: IARC
World Health Organization
World Health Organization
© WHO 2015. All rights reserved

Figure 1: Estimated stomach cancer incidence worldwide in 2012 in men
(Source Globocan 2012 [1])

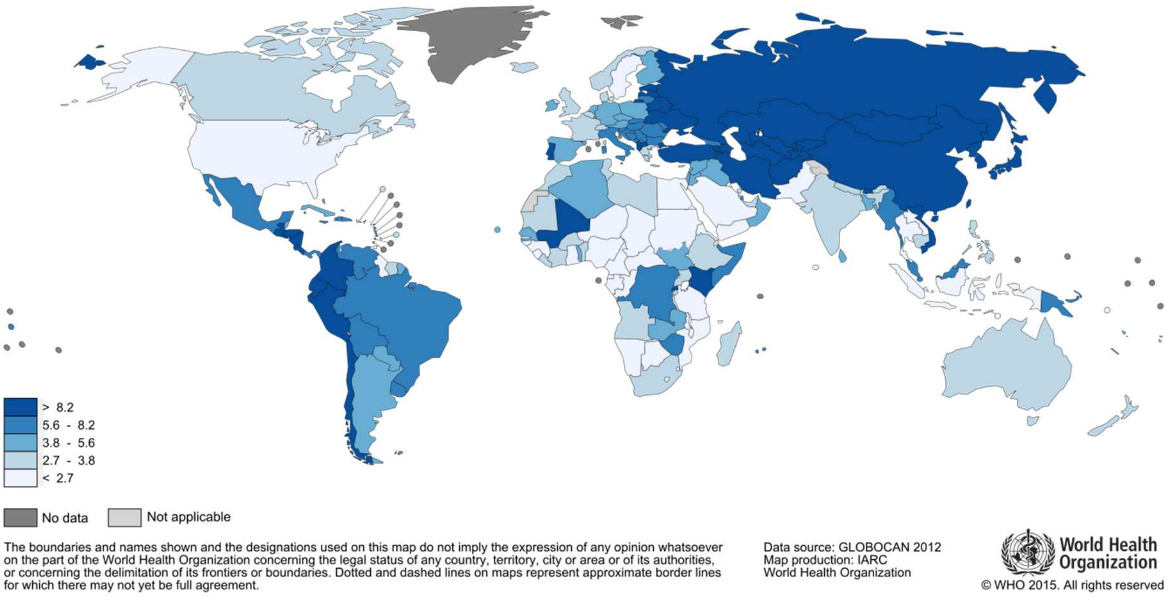

> 8.2
5.6 - 8.2
3.8 - 5.6
2.7 - 3.8
< 2.7
No data
Not applicable
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.
Data source: GLOBOCAN 2012
Map production: IARC
World Health Organization
World Health Organization
© WHO 2015. All rights reserved

Figure 2: Estimated stomach cancer incidence worldwide in 2012 in women
(Source Globocan 2012 [1])

Over the recent decades, gastric cancer mortality rates steadily declined worldwide [2] and although the reasons for these declines are not fully understood. Almost certainly, the reasons of these declines include a more varied and affluent diet and a better food, including refrigeration as well as the control of *helicobacter pylori (HP)* infection, a key determinant of gastric cancer [3].

Other factors appear to have a role in the etiology of the disease [4] some are not modifiable such as age and sex whereas others, such as smoking and *HP* infection, potentially are and prevention can be made in that direction.

Regarding the effect of diet, in 2007, the World Cancer Research Fund and the American Institute for Cancer Research (WCRF/AICR) concluded that non-starchy vegetables, allium vegetables and fruit probably protect against cancer, and that salt including salted and salty food increases the risk of gastric cancer [5]. A meta-analysis of studies on dietary patterns (defined *a priori*¸ *i.e.*, on the basis of specific data under consideration) found an about two-fold difference in gastric cancer risk between a dietary pattern characterized rich in fruit and vegetables and an another one characterized rich in starchy foods, meat and fats [6]. The detrimental effects of processed meat and grilled animal has also been suggested by the WCRF/AICR [5].

Tobacco smoking is an established risk factor of many cancers and chronic diseases. In particular, in the last years, several studies confirmed a positive association between smoking status and gastric cancer [7]. Two recent meta-analyses considering 32 cohort studies [8] and 46 case-control studies [9] showed a significant increasing risk of gastric cancer of 53% and 69%, respectively, in current smokers compared to never smokers. A recent study estimated the worldwide burden of gastric cancer attributable to tobacco smoking in 2012, and found a median of 19.5% for men and 3.0% for women [10].

Risk of gastric cancer was also reported to increase with increasing doses and duration of cigarette smoking. The meta-analysis of cohort studies [8] showed an increasing trend in risk with a relative risk (RR) varying from 1.3 for the lowest doses, to 1.7 for 30 smoking cigarettes per day. A significant trend in gastric cancer risk with increasing duration was reported in the European Investigation into Cancer and Nutrition (EPIC) [11] and the Multiethnic Cohort (MEC) study [12]. Similarly, a recent meta-analysis considering 10 studies on gastric cardia adenocarcinoma reported an over two-fold risk for smokers of more than 40 years compared to never smokers [13].

Risk has been generally found to be lower in former smokers compared to current smokers and seems to decrease with increasing years since stopping smoking, although none found statistically significant dose-response relationships [7, 11, 14-17].

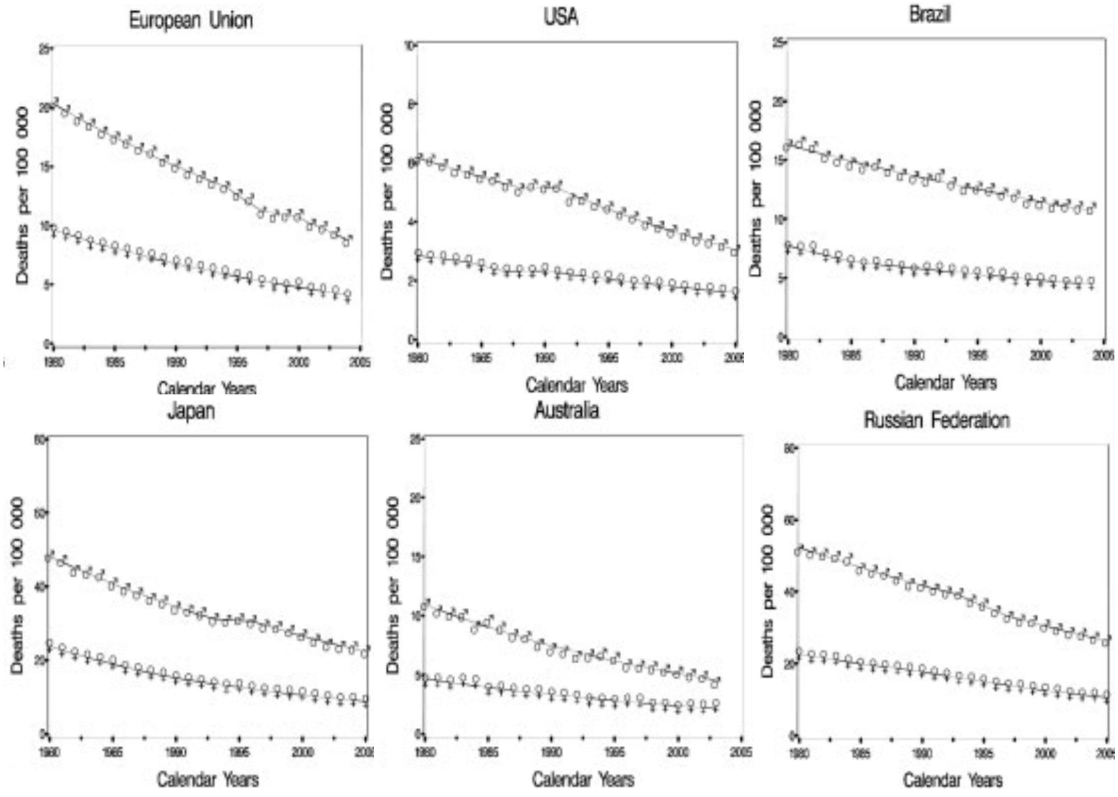


Figure 3: Joinpoint Analysis for gastric cancer in selected countries at all ages (men and women), 1980-2005 (when available). Men ♂—♂ ; Women ♀—♀
(Source Bertuccio et al. International Journal of Cancer 2009 [2])

Various consortia of epidemiological studies have been established during the last two decades, to pool and analyze data on risk factors for breast, ovarian, head and neck, pancreatic, thyroid and other neoplasms. These allowed to identify, and to better quantify the role of important risk factors for various cancers [18, 19]. Because of larger sample sizes, it also offers to consider uncommon exposure, rare diseases, and lead subgroup analyses with greater statistical power than is possible in individual studies (ref Smith-Warner). However, a similar project has not yet been conducted for gastric cancer. A lot of large case-control studies on this neoplasm have been conducted over the years, and a concerted strategy for the joint analysis of these investigations may allow new insights on gastric cancer etiology.

The aim of my PhD was to work on a consortium of case-control studies on gastric cancer, helping in the management of the project (creation of the core variables, harmonization of the datasets and participation of international meeting) and conducting analyses on risk factors using adequate statistical approaches.

During my first year of PhD, I studied the two-stage method. This method is often used to analyze meta-analysis and could be applicable in a case of pooled case-control analysis [20, 21]. The first step of the method consists in calculate adjusted study-specific odds ratios (OR) in order to overcome differences across studies in terms of design or population. The second step consists in summarize these study-specific risks using meta-analytic methods which take into account the heterogeneity across studies.

During my second year of PhD, I studied different approaches to analyze continuous risk variables. The standard approach is to categorize the exposure but it exists other methods avoiding cutpoints including non parametrical methods (such as generalized additive models) or parametrical methods using for example fractional polynomial regressions or regression splines.

During my third year of PhD, I focused my research on a way to adapt these latest methods to the analysis of pooled case-control studies. In particular I chose to use factional polynomials in a two-stage method due to their simple interpretation and also because their estimates can be easily pooled through a two-stage analysis.

II. STATISTICAL APPROACH FOR STUDYING CONTINUOUS RISK FACTORS IN A CONSORTIUM OF CASE-CONTROL STUDIES

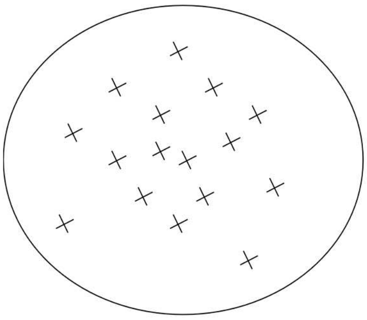
In these pooled data analyses, we will study the effect of risk factors on the occurrence of gastric cancer. The standard approach for the statistical analysis is to conduct aggregate analyses, using data as a unique dataset and to estimate odds ratio and corresponding 95 % confidence interval using multivariable logistic regression adjusted for the variable identifying the study and other potential confounding variables.

However, in this case, some problems arise that the aggregate analyses cannot take into account.

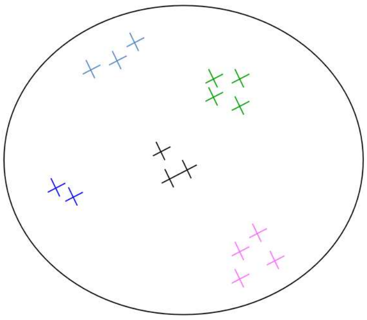
In fact, there is a problem of correlated or clustered data:

- The binary outcome variable is observed in a **group** or a **cluster**
- Each members of the group is correlated with the other members of the same group
- Each group could have its own specific variables
- The variability across groups is high

Not grouped data



Grouped data



To take into account this information, one of the approaches that could be used is the **two-stage analysis method**. This method is often used to analyze meta-analysis and could be applicable in a case of pooled case-control analysis [20, 21].

The two-stage method is an approach that consists of estimating the effect of a uniformly-defined exposure variable within each study and then combining these estimates across studies.

1. The two-stage analysis

1.1. The logistic regression model

We consider k studies (k=1,…, K). The first step of the two-stage method is to perform a logistic regression model for each study. The logistic regression describing the effect of the exposure X on the disease, that is characterized by the presence (Y=1) or the absence (Y=0), adjusting for Z, a confounder that may differ across studies. The model is written as,

$$\text{Logit (P(Y = 1 | X, Z)} = \alpha_{jk} + \beta_k X_{ik} + \gamma_k Z_{ik} \qquad (1)$$

where j: stratum of study {j=1,…S_k}

k: the identification of the study {k=1,…,K}

i: individual cases and controls {i=1,…n_{jk}}

α: intercept

β: the parameter estimated for X

γ: the parameter estimated for Z

The exposure X is uniformly defined across studies. However, the confounders Z_k may be specific to a particular study and may vary in definition across studies. To simplify we assume only one confounder per study but this is easily generalized to more.

The logistic regression can be performed with the SAS software (SAS Institute Inc, Cary, NC) using the PROC LOGISTIC procedure.

1.2. Test of homogeneity

To choose the adequate model of the second step of the two-stage model, a test of homogeneity between the studies is required. It allows to evaluate the consistency of exposure effects across the studies.

The hypothesis of the test of homogeneity is:

$$\begin{cases} H_0: \beta_1=\beta_2=\ldots=\beta_K \\ H_1: \text{At least one of the } \beta_k \text{ is different} \end{cases}$$

Under the null hypothesis t the test statistic is defined by:

$$Q = \sum_{k=1}^{K} w_k \left(\hat{\beta}_k - \hat{\beta}\right)^2 \qquad (2)$$

where $\hat{\beta} = \frac{\sum_{k=1}^{K} w_k \hat{\beta}_k}{\sum_{k=1}^{K} w_k}$ and $w_k = \frac{1}{\hat{\sigma}_k^2}$.

In particular, $\hat{\beta}$ is the estimation of the pooled exposure log-odds ratio and σ_k^2 represents the within-study variation of the β_k.

The Q test statistic follows a Chi^2 distribution with k-1 degrees of freedom, $Q \sim \chi_{k-1}^2$.

When there is homogeneity of exposure effects across studies, we can assume that the variance across studies $\hat{\theta}^2$ is null ($\hat{\theta}^2 = 0$) and exposure effects can be estimated through **a fixed effect model**.

A high value of Q indicates a high variability across studies and when the null hypothesis is rejected, it means that exposure effects are not homogeneous between studies; in this case, the pooled-exposure effect is generally estimated using **a random effect model**.

1.3. The fixed effect model

The fixed effect model is generally used when the exposure-effect is not different across studies. The second stage of the model is defined by

$\beta_k = \beta + e_k$ with $\beta_k \sim N(\beta, \sigma_k^2)$ and $e_k \sim N(0, \sigma_k^2)$.

Thus, the pooled-exposure effect is a simple weighted average of the β_k

$\beta = \frac{\sum_{k=1}^{K} w_k \beta_k}{\sum_{k=1}^{K} w_k}$ with weights w_k equal to the inverse of the variance, $w_k = \frac{1}{\mathrm{var}(\beta k)} = \frac{1}{\sigma_k^2}$.

The variance of the pooled effect β is equal to $Var(\beta) = \frac{1}{(\sum_{k=1}^{K} w_k)}$.

The sample estimates of the above quantities are:

$$\hat{\beta} = \frac{\sum_{k=1}^{K} \widehat{w}_k \widehat{\beta}_k}{\sum_{k=1}^{K} \widehat{w}_k} \qquad \text{with} \quad \widehat{w}_k = \frac{1}{\widehat{\sigma}_k^2} \tag{3}$$

1.4. The random effect model

The random effect model is generally used when the exposure effect is different across studies. This supposes that the pooled-exposure effect β_k varies across studies around the real parameter β with a variance θ^2 according to the second-stage model:

$$\beta_k = \beta + b_k + e_k \qquad \text{with } \beta_k \sim N(\beta, \sigma_k^2 + \theta^2) \tag{4}$$

where β is the pooled exposure log-odds ratio

b_k are random effects with $b_k \sim$ N (0, θ^2)

θ^2represents the variability of the study-specific exposure effects β_k about the population mean β.

e_k are independent errors with $e_k \sim N(0, \sigma_k^2)$

σ_k^2 represents the within-study variation of the β_k

The estimation of the pooled-exposure effect β is the weighted average of the β_k , weighted by the inverse marginal variance of the $\hat{\beta}_k$ as follows:

$$\hat{\beta} = \frac{\sum_{k=1}^{K} \widehat{w}_k \widehat{\beta}_k}{\sum_{k=1}^{K} \widehat{w}_k} \quad \text{with} \quad \widehat{w}_k = \frac{1}{\widehat{\sigma}_k^2 + \widehat{\theta}^2}$$

(5)

and $\mathrm{var}(\hat{\beta}) = \frac{1}{\sum_{k=1}^{K} w_k}$

To compute the estimation of the pooled-exposure effect, an estimate of the random effects variance is required. Two methods are frequently used: the **moment estimation** and the **pseudo-maximum likelihood**.

The variance θ^2 of a random effects model is a measure of the heterogeneity across studies (a fix effects model is a particular case where $\theta^2 = 0$)

- **The moment estimation of θ²**

The moment method compares the observed and expected values of the Q statistic [20].

$$Q = \sum_{k=1}^{K} w_k \ (\hat{\beta}_k - \hat{\beta})^2$$

Thus, $E(Q) = E\left[\sum_{k=1}^{K} w_k \ (\hat{\beta}_k - \hat{\beta})^2\right]$

$$= \sum_{k=1}^{K} w_k \, E(\hat{\beta}_\mathrm{k}^2) - \sum_{k=1}^{K} w_k \, E(\hat{\beta}^2)$$

$$= \mathrm{k} - 1 + \theta^2 \ \left[\sum_{k=1}^{K} w_k - \frac{\sum_{k=1}^{K} w_k^2}{\sum_{k=1}^{K} w_k}\right]$$

$= Q$

The estimation of θ^2 can be derived resolving the equation above and

$$\hat{\theta}^2 = \begin{cases} \hat{\theta} \ if \ \hat{\theta} > 0 \\ 0 \ \ if \ \hat{\theta} \leq 0 \end{cases}$$

Where $$\hat{\theta} = \frac{Q - (k-1)}{\sum_{k=1}^{K} w_k - \left(\sum_{k=1}^{K} w_k^2 \big/ \sum_{k=1}^{K} w_k\right)} \quad (6)$$

Where Q and w_k are described above. This estimator is unbiased and non-iterative.

Since the moment estimation of θ^2 and the calculation of Q and $\hat{\beta}$ require matrix calculations, the SAS procedure PROC IML can be used.

- **The pseudo-maximum likelihood estimation of θ²**

The approach used to estimate the variance θ² by the maximum likelihood is the restricted maximum likelihood (REML) method.

The estimator REML of θ_{r+1}^2 is:

$$\hat{\theta}_{(r+1)}^2 = \theta_r^2 \left(\frac{\sum_{k=1}^{K}\left(\hat{\beta}_k - \hat{\beta}_{(\mathrm{r})}\right)^2\left(\hat{\sigma}_k^2 + \hat{\theta}_{(\mathrm{r})}^2\right)^{-2}}{\sum_{k=1}^{K}\left(\hat{\sigma}_k^2 + \hat{\theta}_{(\mathrm{r})}^2\right)^{-1}}\right) \quad (7)$$

Where $\hat{\beta}_{(\mathrm{r})}$ is recomputed at the *r*th iteration from (5)

$\hat{\theta}_{(0)}^2 = \frac{1}{K}\sum_{k=1}^{K}[\left(\hat{\beta}_k - \hat{\beta}\right)^2 - \hat{\sigma}_k^2]$ is an initial estimate of θ²

$\hat{\beta}_{(0)} = \frac{\sum_{\mathrm{k=1}}^{\mathrm{K}} \hat{\beta}_k/\hat{\sigma}_k^2}{\sum_{\mathrm{k=1}}^{\mathrm{K}} 1/\hat{\sigma}_k^2}$, the weighted average of the study-specific $\hat{\beta}_k$ is an initial estimate of β

To obtain $\hat{\beta}$, on first computes $\hat{\beta}_{(0)}$ and $\hat{\theta}_{(0)}^2$ and then iterates between computing $\hat{\theta}_{(r)}^2$ and $\hat{\beta}_{(\mathrm{r})}$ until convergence.

These estimations can be performed with the SAS procedure PROC MIXED.

Among all risk factors considered in the StoP project, some variables can be continuous. The standard approach to model continuous risk variables is 1) to categorize the exposure into two or more categories, creating dummy and then calculating the effects using one category as reference group or 2) using a linear model to describe the relation between exposure and effect. These methods present the advantage of a simple epidemiologic interpretation but include a loss of statistical efficiency and important errors in particular if the measured relation is not linear.[22] The approaches to overcome limitations related to these methods can be non-parametric (such as generalized additive models) or parametric, *i.e.* fractional polynomial regressions or regression splines.

During the third year of PhD, I decided to put an emphasis on the analysis of the effect exposure variable measured on a continuous scale through a two-stage analysis using as reference a methods developed on meta-analysis of published data [23].

I focused this analysis on fractional polynomials because in the epidemiological context it is a simple but flexible approach. I contrasted it with the traditional analysis using categories.

2. Fractional polynomials through a two-stage analysis: first step

We consider k studies (k=1,…, K). The first step of the two-stage method is to perform a logistic regression using fractional polynomials for each study.

Fractional polynomials were developed by Royston and Altman [24] to look for nonlinearity. They are an extension of polynomials where the exponents can be negatives and/or integer and are usually chosen from the predefined set P={-2;-1;-0.5;0;0.5;1;2;3}.

2.1. First order fractional polynomials

The polynomial is characterized by

$$\boxed{\mathrm{FP}_{1\mathrm{k}} = \alpha_k + \beta_k {X_k}^p + \gamma_k Z_k}$$

where X the exposition variable

Z the confounders

k: the identification of the study {k=1,…, K}

α: intercept

γ: the parameter estimated for Z

β: the parameter estimated for X

p the power term, p ∈ P={-2;-1;-0.5;0;0.5;1;2;3} with

$$\mathrm{X}^{\mathrm{p}} = \begin{cases} \mathrm{X}^{\mathrm{p}} & \text{if } \mathrm{p} \neq 0 \\ \ln \mathrm{X} & \text{if } \mathrm{p} = 0 \end{cases}$$

For example,

p_1=2 the model is $FP_1 = \beta_0 + \beta_1 X^2$

p_1= 0 the model is replace by $FP_1 = \beta_0 + \beta_1 \ln X$

Hence, there are 8 different first-order fractional polynomials (FP1) models.

For p=1, the linear model is generated. From the predefined set P, some important transformations are generated such as the inverse (p=-1), the squared root (p=0.5), the logarithm (p=0) and the quadratic (p=2) transformation.

2.2. Second order polynomials

Second order polynomials are defined as follows

$$\boxed{\mathrm{FP}_{\mathrm{mk}} = \alpha_k + \sum_{j=1}^{m} \beta_{jk} {X_k}^{p_j} + \gamma_k Z_k}$$

where m the degree of the fractional polynomial j=1…m

k: the identification of the study {k=1,…, K}

α: intercept

β: the parameter estimated for X

X the exposition variable

γ: the parameter estimated for Z

Z the confounders

p_j the power terms, $p_j \in P=\{-2;-1;-0.5;0;0.5;1;2;3\}$ with

$$\mathrm{X}^{\mathrm{pj}} = \begin{cases} \mathrm{X}^{\mathrm{p\,j}} & \text{if } \mathrm{p} \neq 0 \\ \ln \mathrm{X} & \text{if } \mathrm{p} = 0 \\ \mathrm{X}^{\mathrm{p\,j}} \ln \mathrm{X} & \text{if } \mathrm{p_j} = \mathrm{p_{j-1}} \end{cases}$$

For example,

if m=2, p_1=1 and p_2=2 the model is FP2= $\beta_0 + \beta_1 X + \beta_2 X^2$

if m=2 and p_1= p_2 the model is replace by FP_2= $\beta_0 + \beta_1 X^p + \beta_2 (X^p \ln X)$

Hence, there are 36 different second order fractional polynomials (FP2) models. In practice, it has been observed that it is rarely necessary to consider degrees higher than 2 so we considered the second-order fractional polynomial for the rest of this report.

Second order fractional polynomials can be monotonic or unimodal (i.e. with a maximum or a minimum point for some positive value of X). The value of X for the minimum or maximum point of the function, can be derived from the formulae given in the Table 1.

Table 1. Minimum or maximum point for second order fractional polynomials based on the power (p1 ; p2) values and the model estimates (β1 ; β2). r= - β1/β2.

Powers value	**p1=0**	**p1≠0**
p2=0 (p2≠p1)	$\nexists$	$(rp_1)^{-1/p_1}$
p2≠0 (p2≠p1)	$(r/p_2)^{1/p_2}$	$(rp_1/p_2)^{1/(p_2-p_1)}$
p2=p1	$\exp(r/2)$	$\exp(r-1/p_1)$

Models generated with the second-order fractional polynomial technical are ranging from U-shaped to J-shaped relationships

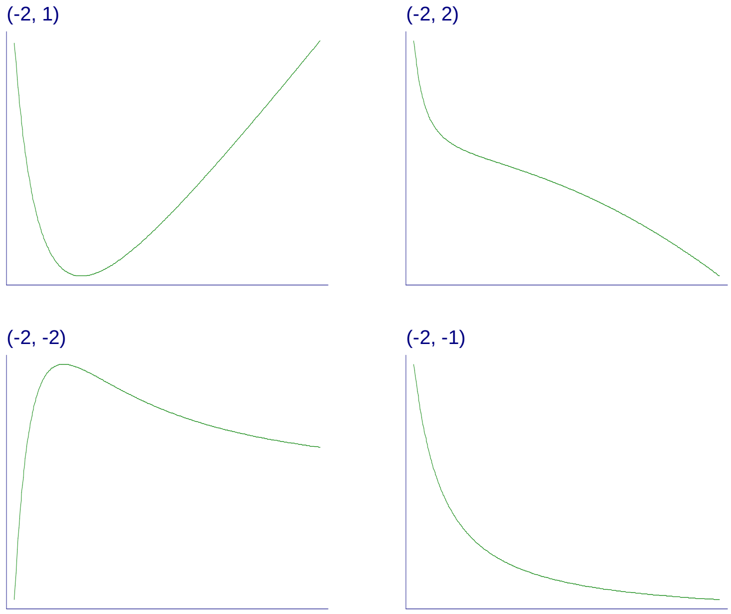


Figure 4: Some examples of curve shape with second-degree fraction polynomials with p_1=-2 and p_2 varying from -2 to 1.

The confounders Z_k may be specific to a particular study and may vary in definition across studies. To simplify, then we assume only one confounder per study but this is easily generalized to more.

3. Fractional polynomials through a two-stage analysis: second step

For each value of the power term (or couple of power terms for the second-order fractional polynomials), the second stage of the model is performed.

The pooled dose-response relationship is estimated accordingly to the bivariate random-effects model:

$$\begin{pmatrix}\beta_{1k}\\ \beta_{2k}\end{pmatrix} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} + \begin{pmatrix}b_{1k}\\ b_{2k}\end{pmatrix} + \begin{pmatrix}e_{1k}\\ e_{2k}\end{pmatrix}$$

$$\begin{pmatrix}\beta_{1k}\\ \beta_{2k}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \sigma_k^2 + \theta^2\right)$$

Where μ_1 and μ_2 are the pooled trend component

σ_k^2 is the within study variance matrix of the kth study

θ^2 is the between-studies variance matrix that has to be estimated:

$$\theta^2 \ = \begin{pmatrix} var(\mu_1) & cov(\mu_1, \mu_2) \\ cov(\mu_1, \mu_2) & var(\mu_2) \end{pmatrix}$$

e_1 and e_2 are independent error with $\begin{pmatrix}e_{1k}\\ e_{2k}\end{pmatrix} \sim N(0, \sigma_k^2)$

The estimate μ=(μ_1, μ_2) of the trend component μ_1 and μ_2 could be obtained using restricted maximum likelihood (REML) or maximum likelihood (ML) estimation [25].

The second stage model is fitted to the data considering each combination of the power terms. The best model, denoted by the optimal power combination (p_1, p_2) is defined as the one minimizing the deviance or the Akaike Information Criterion (AIC), a penalized likelihood which takes into account the number of parameter .

Finally the expected log(OR) at exposure level *x* can be estimated from the optimal model as

$$\log(OR) = \mu_1 x^{p_1} + \mu_2 x^{p_2}$$

with a 95% confidence interval equal to $\log(OR) \pm 1.96\sqrt{t_X cov(\mu) X}$

where $X = (x^{p_1}, x^{p_2})$ the power transformation for the assigned dose *x*

t$_X$ is the transpose of X

$cov(\mu) = (\sum_{k=1}^{K} (v_k + T)^{-1})^{-1}$

where T is the estimated of θ^2, the between-studies variance matrix

vj is the estimated of σ_j^2, the within study variance matrix of the jth study

These estimations can be performed with the SAS procedure PROC MIXED.

As a strategy for selecting the best model the following selection procedure is proposed [26, 27]

1. Overall association

   Test the best FP2 model against the null model using 4 degrees of freedom. If the test is not significant, it means that the effect of the exposure is not significant. The analysis can stop at this point.

2. Evidence for non-linearity

   Test the best FP2 model against the straight line using 3 degrees of freedom. If the test is not significant, it means that the relationship between the risk of disease and the exposure, is linear. The final model is a straight line.

3. Test the best FP2 against the best FP1 using 2 degrees of freedom. If the test is not significant, the final model is FP1, otherwise the final model is FP2.

We elaborated a SAS macro for both first-order and second order polynomial through a two-stage analysis.

III. DESCRIPTION OF THE STOP PROJECT

The Stomach cancer Pooling (StoP) Project is an international epidemiological consortium. The inclusion criteria for study participation are: a case-control study design (including nested case-control analyses derived from cohort study) and an inclusion of at least 80 cases of gastric cancer (including both cardia and non-cardia location).

To date, 34 studies from 14 countries agreed to participate (2 from Brazil, 1 from Canada, 3 from Mexico, 2 from USA, from Greece, 4 from Italy, 1 from Portugal, 1 from Russia, 2 from Spain, 3 from Sweden, 1 from Latvia, 4 from China, 3 from Iran, 3 from Japan), for a total of around 13,000 cases and 31,500 controls, and contacts are ongoing with investigators involved in 6 other studies (1 from Finland, 1 from Poland and 3 from the USA), for potential inclusion of these investigations during the next months (summary information on preliminary data available Table 2).

1. Characteristics of each study

Table 1 describe the main characteristics of each study.

The study 7 from Canada (Principal investigator, PI: K. Johnson) [28] was conducted in 8 provinces (British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Prince Edward Island, Nova Scotia and Newfoundland) between 1994 and 1997. It included 1182 cases (379 women and 803 men) and 5039 controls (2492 women and 2547 women) matched with cases on age and sex.

The study 14 MSKCC (Memorial Sloan Kettering Cancer Center) (PI: ZF Zhang) [29] was conducted in New-York between November 1992 and November 1994. It included 134 incident cases (99 men, 35 women) and 132 controls (62 men, 70 women) classified cancer-free after endoscopic examination in the same endoscopy unit of the cases at the MSKCC.

The study 16 from New-York, USA (PI: J Muscat) (unpublished data) was conducted between 1980 and 1990 on 87 (78 men, 9 women) incident cases and 261 hospital-based controls matched to cases on age and sex.

The study 6 from the greater Athens area, Greece (PI: D Trichopoulos) [30] was conducted between May 1981 and June 1984, on 110 incident cases (57 men, 53 women, mean age 64.5, range 34-85) with histologically confirmed adenocarcinoma of stomach and 100 hospital-based controls (49 men, 51 women, mean age 59.8, range 34-84).

The study 1 from the greater Milan area, Italy (PI: C La Vecchia) [31] was conducted between 1985 and 1997 on 769 incident, histologically confirmed cases of stomach cancer (469 men, 300 women, median age 61 years, range 19-80 years) and 2081 controls (1220 men, 861 women, median age 55 years, range 19-80 years). Controls were subjects admitted to the same network of hospitals as cases.

The study 3 from the greater Milan area, Italy (PI: E Negri) [32] was conducted between 1997 and 2007 on 230 incident cases of gastric cancer (143 men, 87 women; median age 63 years, range 22–80 years) and 547 controls (286 men, 261 women, median age 63 years, range 22–80 years), frequency-matched with cases by age and sex. Controls were subjects admitted to the same network of hospitals as cases.

The study 4 from Rome, Italy (PI: S Boccia) [33] was conducted from November 1999 to February 2005 on 76 cases (37 men, 39 women, mean age 66.1, range 32-89) and 260 control controls (145 men, 115 women, mean age 63.8, range 30-91). Controls were selected from the same hospital as cases and matched to cases on age and sex.

The study 5 from 4 areas in Italy (PI: D Palli) [34] involved 2 areas with high death rates for gastric cancer (1: Forlì, Cremona and Imola and 2: Florence and Siena) and 2 areas with low death rates for gastric cancer (3: Genoa and 4: Cagliari). It included 1229 cases (640 men, 376 women, median age 65) and 1159 controls (705 men, 454 women) matched with cases on age and sex.

The study 17 from Porto, Portugal (PI: N Lunet) [35] was conducted between 1999 and 2006 and included 568 cases (353 men, 215 women, age range 18-92) and 1585 controls. Cases and controls were frequency matched on age and sex. Controls were part of a representative sample of the adult population of Porto.

The study 9 from Moscow, Russia (PI: D Zaridze) [36] was conducted between 1996 and 1997 on 448 cases (248 men, 200 women) and 610 hospital-based controls (292 men, 318 women).

The study 21from Spain (PI: N Aragones) (unpublished data) was conducted in Asturias, Barcelona, Cantabria, Granada, Huelva, Leòn, Madrid, Murcia, Navarra and Valencia, between 2008 and 2012. Around 400 incident cases and 1800 controls were included and were matched on age, gender and recruitment area. Controls were selected in the general population residing in the catchment areas of the hospitals where cases were recruited. The final dataset of the study is under preparation, therefore the precise number of included subjects is not yet available.

The other study from Spain, study 23, (PI: J. Vioque Lopez) [37] was conducted between January 1995 and March 1999 in 9 hospitals in Alicante and Valencia on 399 incident histological cases (265 men, 134 women) and 455 hospital-based controls (285 men, 170 women) frequency matched by sex, axe and province of residence.

The study 22 from Sweden (PI: O Nyren) [38] was conducted from 1989 to 1995 on 514 cases (348 men, 166 women) and 1164 controls (779 men, 385 women). Controls were randomly selected from population registers and were frequency matched to cases on age and sex.

The two other studies 18 and 20 from Sweden (PI: N Orsini) (unpublished data) was conducted in two counties of central Sweden, Vastmanland and Uppsala on women only for the first one and Vastmanland and Orebro on men only for the second one. They are nested case-control studies, derived from the Swedish mammography cohort (93 cases and 372 controls) and the cohort of Swedish men studies (176 cases and 704 controls).

The study 2 from Harbin, China (PI: J Hu) [39] was conducted from March 1987 to May 1989 on 266 newly diagnosed and histologically confirmed stomach cancer patients (206 Men and 60 women, median age 57 years, range 23-80). Controls were 533 patients (412 men, 121 women, median age 57, range 22-79) admitted to the same hospitals for non-neoplastic and non-gastric diseases. Cases were not individually matched to controls, but were well comparable by age and sex.

The study 8 from Taixing, China (PI: L Mu) [40] was conducted in 2000 and included 206 cases (168 men, 68 women, range 30-82) and 415 controls (287 men, 128 women, range 21-84) randomly selected in the general population. Cases and controls were frequency matched by age and sex.

The study 12 from Shangai and Qingdao, China (PIs: Yu and ZF Zhang) [41] was conducted between 1991 and 1993 on 951 incident primary stomach cancer cases (621 men, 330 women, mean age 62.5) and 951 controls (621 men, 330 women, mean age 62.1). Controls were selected from the general population in the same street or community of the cases and matched to cases on age and sex.

The study 13from Yangzhong, in Jiangsu province in the southeast of China (PI: ZF Zhang) [42] was conducted from January 1995 to June 1995. It included 133 newly diagnosed cases (93 men, 40 women) and 433 controls (214 men, 219 women) selected from a name list of residents in Yangzhong.

The three studies 10, 11 and 19 were conducted in Ardabil, Iran (PI: R Malekzadeh).

The first one [43] was conducted in 1999 for cases and from 2003 to 2005 for controls. It included 217 cases (151 men, 66 women, mean age 65.4) recruited through the Ardabil cancer registry, and 394 controls (265 men, 129 women, mean age 64.3) randomly selected form the annual household survey of the health department.

The second one [44] was conducted from August 2005 to August 2007 on 286 cases (210 men, 76 women, mean age 66.3) and 304 controls (217 men, 87 women, mean age 62.9). Controls were not individually matched to cases and were selected to be representative of the Ardabil population aged other 40 years living in rural and urban areas.

The third one [45] included 119 cases (86 men, 33 women, mean age 65.0) and 119 controls, selected from dyspeptic patients and matched to cases on sex, age and centre.

The study 15 from Aichi, Japan (PI: K Matsuo) (unpublished data) was conducted between 2001 and 2005. Cases were selected from the HERPACC-II (Hospital-based Epidemiologic Program at Aichi Cancer Canter-II) study which enrolled all first outpatients visit aged 20-79, regardless of cancer status. There were 1250 cases (882 men, 368 women). The 3911 controls were also recruited from the participants of HERPACC, and were subjects diagnosed as not having cancer within 1-year from random sampling. Cases and controls were individually matched on age and sex.

The study 24 from Japan (PI: H. Ito and K. Matsuo) [46] included 2552 cases and 5138 hospital controls from the HERPACC-I (1988-2001). Controls were randomly selected and individually age-, sex- and enrolment year-matched to cases with a 1: 2~3 case-control ratio.

The study 25, 26 and 27 (PI: L. Lopez-Carillo and R.U. Hernandez-Ramirez) [47]

The first study [48] was conducted in Mexico city between 1989 and 1990 on 220 histologically confirmed newly diagnosed cases (122 men and 98 women) and 752 controls (296 men and 456 women). Controls were frequency matched by age +/- 5 years and recruited from residents of the Mexico city metropolitan area

The second study [47] was conducted in Mexico city, Merida and Puebla in Mexico between 1994 and 1996 on 324 cases (133 men and 101 women) Histologically confirmed incident adenocarcinomas of the stomach and 468 controls (266 men and 202 women) matched to case by age (±5 years), sex and city of residence. Controls were recruited in the same hospital as cases.

The third study [49] was conducted in Mexico city between 2004 and 2005 on 248 cases (134 men, 114 women) and 478 controls (258 men and 220 women) recruited from the general population.

Two studies 28 and 29, from Brazil were conducted between 1991-1994 (PIs: S. Tsugane and G.S. Hamada), one was conducted on Japanese Bresilian [50, including 96 cases and 192 age-, sex-, and race-matched controls, and one was conducted on Non-Japanese Brazilian {Nishimoto, 2002 #541] including 236 cases and 236 age-, sex-, and race-matched controls.

The study 30, from Japan [51] was conducted from 1998 to 2002 (PI: S. Tsugane) in 4 hospitals in Nagano and included 153 cases and 301 age-, and sex-matched controls (participants of health check-up).

The study 31, from Latvia began the recruitment in 2007 and is still ongoing (PI: M. Lja and E. Gasenko) (unpublished data) and project to include 400 cases and 1100 controls.

Table 2. Characteristics of the 31 studies included in the StoP project[1]

City/Region, Country	Investigator (recruitment period)	N cases	N controls	Dataset avai-lable	Dataset harmo-nized
America - 6 studies		**2424**	**7542**		
Sao Paulo, Brazil 1	S. Tsugane (1991-1994)	93	186	-	-
Sao Paulo, Brazil 2	S. Tsugane (1991-1994)	226	226	-	-
8 provinces, Canada	K. Johnson, J. Hu (1994-1997)	1182	5039	X	X
Mexico 1, Mexico	L. Lopez-Carillo (2004-2005)	248	478	X	-
Mexico 2, Mexico	L. Lopez-Carillo (1989-1990)	220	752	X	-
Mexico 3, Mexico	L. Lopez-Carillo (1994-1996)	234	468	X	-
New York, MSKCC, USA 1	ZF. Zhang (1992-1994)	134	132	X	X
New York, USA 2	J. Muscat (1980-1990)	87	261	X	X
Europe - 13 studies		**5102**	**11451**		
Athens, Greece	D. Trichopoulos (1981-1984)	110	100	X	X
Greater Milan, Italy 1	C. La Vecchia (1985-1997)	769	2081	X	X
Greater Milan, Italy 2	E. Negri (1997-2007)	230	547	X	X
Roma, Italy 3	S. Boccia (2006 ongoing)	164	444	X	X
4 areas, Italy 4	D. Palli (1985-1987)	1016	1159	X	X
Porto, Portugal	N. Lunet (1999-2006)	568	1585	X	X
Moscow, Russia	D. Zaridze (1996-1997)	448	610	X	X
10 provinces, Spain 1	N. Aragones (2008-2012)	400	1800	X	X
South East, Spain 2	J. Vioque (1995-1999)	399	455	X	X
5 counties, Sweden 3	W. Ye (1989-1995)	514	1164	X	X
2 counties (men), Sweden 1	N. Orsini (1997-1998)	93	372	X	X
2 counties (women), Sweden 2	N. Orsini (1997-1998)	176	704	X	X
Latvia	M. Leja (2007-ongoing)	215	430	-	-
Asia - 10 studies		**6133**	**12499**		
Harbin, China 1	J. Hu (1987-1989)	266	533	X	X
Taixing, Jiangsu, China 2	L. Mu (2000)	206	415	X	X
Shangai, Qingdao, China 3	Y. GuoPei (1991-1993)	951	951	X	X
Yangzhong, China 4	ZF. Zhang (1995)	133	433	X	X
Ardabil, Iran 1	R. Malekzadeh (1999)	217	394	X	X
Ardabil, Iran 2	R. Malekzadeh (2005-2007)	286	304	X	X
Ardabil, Iran 3	R. Malekzadeh	119	119	X	X
Aichi, Japan 1	K. Matsuo (2001-2005)	1250	3911	X	X
Nagoya, Japan 2	H. Ito (1988-2001)	2552	5138	X	-
Nagano, Japan 3	S. Tsugane (1998-2002)	153	301	-	-
TOTAL - 31 studies		**13659**	**31492**		

[1] List at October 2015.

2. First main task of the project: harmonization of datasets

The first task consists on merging the different datasets, each one with its specific variables as well as with different variable names, format and codes, in order to create a single uniform dataset for pooled data analyses. During my first year, I participated on it. For that, we collected the variables available for each study and we divided them in several main topics as listed in Table 2. For each study, we created a codebook reporting which variables are present in each study, their names and their codes. Afterwards, we standardized the formats of variables in order to make them uniform between datasets and to make them available for pooled data analyses.

We began the work on harmonization on 9 groups of variables (cards). These variables were selected among those of first interest for analyses and those required in most analyses particularly for adjustment. The harmonisation of datasets is still ongoing.

IV. APPLICATION ON THE RELATION BETWEEN CIGARETTE SMOKING AND GASTRIC CANCER RISK

1. Studies included in the analysis

We analyzed data from 21 case-control studies of the StoP project, on 10040 cases (6624 men, 3414 women) and 25602 controls (15,305 men, 10,297 women) from China (3 studies), Iran (2 studies) Japan, Canada, USA (2 studies), Italy (4 studies), Greece, Russia, Portugal, Spain (2 studies), and Sweden (3 studies).

Table 3 shows the distribution of cases and controls by study, sex, age and other potential and confounding factors. The proportion of men was slightly higher in cases (66.0%) than in controls (59.8%). Cases were somewhat older and have a social class lower than controls. They reported more frequently a history of stomach cancer in first degree relatives, consumed less vegetables and fruit and declared to drink more alcohol than controls.

Table 3. Distribution of 10 040 cases of pancreatic cancer and 25602 controls according to study center, sex, age, and other selected covariates. Stomach cancer pooling (StoP) consortium.

	Cases		**Controls**	
	N	**%**	**N**	**%**
Total	10040		25602	
Study center				
Asia	*2946*	*29.3*	*5684*	*22.2*
02. China 1 (Hu)	266	2.6	533	2.1
08. China 2 (Mu)	206	2.1	415	1.6
12. China 3 (Zhang-Yu)	711	7.1	711	2.8
10. Iran 1 (Malekzadeh)	217	2.2	394	1.5
11. Iran 2 (Malekzadeh)	286	2.8	304	1.2
15. Japan (Matsuo)	1260	12.5	3327	13.0
North America	*2014*	*20.6*	*7253*	*28.3*
07. Canada (Johnson)	1182	11.8	5039	19.7
14. USA (Zhang)	132	1.3	132	0.5
16. USA (Muscat)	700	7.0	2082	8.1
Europe	*5080*	*50.6*	*12665*	*49.5*
01. Italy 1 (La Vecchia)	769	7.7	2081	8.1
03. Italy 2 (Negri)	230	2.3	547	2.1

04. Italy 3 (Boccia)	161	1.6	444	1.7
05. Italy 4 (Palli)	1016	10.1	1159	4.5
06. Greece (Trichopoulos)	110	1.1	100	0.4
09. Russia (Zaridze)	450	4.5	611	2.4
17. Portugal (Lunet)	692	6.9	1667	6.5
21. Spain 1 (Aragones-Martin)	441	4.4	3441	13.4
23. Spain 2 (Vioque-Navarrete-Munoz)	401	4.0	455	1.8
18. Sweden 1 (Wolk-Orsini)	88	0.9	352	1.4
20. Sweden 2 (Wolk-Orsini)	161	1.6	644	2.5
22. Sweden 3 (Ye)	561	5.6	1164	4.5
Sex				
Male	6624	66.0	15305	59.8
Female	3414	34.0	10297	40.2
Age				
<50	1305	13.0	5208	20.3
50-54	965	9.6	2631	10.3
55-59	1302	13.0	3069	12.0
60-64	1562	15.6	4018	15.7
65-69	1811	18.0	4189	16.4
70-75	1822	18.1	3822	14.9
≥75	1273	12.7	2665	10.4
Social class				
Low	5305	52.8	10354	40.4
Intermediate	2681	26.7	7747	30.3
High	1237	12.3	5385	21.0
Missing	*817*	*8.1*	*2116*	*8.3*
History of stomach cancer in first degree relatives[1]				
No	5014	49.9	12678	49.5
Yes	876	8.7	1271	5.0
Missing	*4150*	*41.4*	*11653*	*45.6*
Vegetables and fruit intake[2]				
Low	3027	30.1	6807	26.6
Intermediate	3102	30.9	7655	29.9
High	2998	29.9	8225	32.1
Missing	*913*	*9.1*	*2915*	*11.4*
Alcohol drinking (gr of alcohol/day)[3]				
Never	2440	24.3	7086	27.7
Low (< =12)	2080	20.7	7257	28.3
Intermediate (>12 and <=47)	2406	24.0	5379	21.0
High (>47)	1134	11.3	2262	8.8
Missing	*1980*	*19.7*	*3618*	*14.1*

[1] No information available for studies China 1 (Hu), Canada (Johnson), China 3 (Zhang-Yu), USA 2 (Muscat), Sweden 1 (Wolk-Orsini) and Sweden 2 (Wolk-Orsini)

[2] No information available for the study USA (Muscat)

[3] Alcohol drinking was not available in category of consumption for the study Iran 2 (Malekzadeh), China 3 (Zhang-Yu), Sweden 3 (Ye)

2. Exposure variable: Cigarette smoking

All studies in this pooled analysis provided information about cigarette smoking status (never, former, and current smoker), number of cigarettes smoked per day, duration of smoking, and time since stopping. Though questions about cigarette smoking were similar across studies, we conducted a careful and detailed examination of the comparability of smoking-related questions to harmonize the data from the multiple studies included in this pooled analysis.

For the present analyses, ever cigarette smokers were defined as participants who had smoked at least 100 cigarettes in their lifetime or more than one cigarette per day for at least 1 year.

For some variable related to the duration of smoking and former smoker status, when the type of smoking could not be deduce (cigarette, pipe or cigars) the data was not considered. However, when the study did not provide information on the type of smoking for the entire smoking variables we considered smoking status (ever, never) as valid for cigarette smoking.

3. Statistical methods

To estimate the association between cigarette smoking and pancreatic cancer risk, we used a two-stage modeling approach [52]. In the first stage, for categorical variables we assessed the association between cigarette smoking and gastric cancer for each study by estimating the odds ratios (ORs) and the corresponding 95% CIs using multivariable unconditional logistic regression models. These models included, when available terms for age (<40, 40-45, 45-50 50–54, 55–59, 60–64, 65–69, 70–74, ≥75 years), sex, education (study-specific low, intermediate, high), race/ethnicity (White, Hispanic/Latino, Black/African american, other), alcohol drinking consumption (Never, low ≤12 gr/day, intermediate >12-≤47 gr/day, high >47 gr/day) and study center for multicentric studies.

For continuous variable, we assessed the estimation of the odds ratios (ORs) and the corresponding 95% CIs using one-order and two-order fractional polynomial models. The best fitting model was define as the one minimizing the deviance.

Using a macro program we developed on SAS software (SAS Institute Inc, Cary, NC) (See Supplements), 8 first-order and 36 second-order fractional polynomial models were generated with the power vector P={-2;-1;-0.5;0;0.5;1;2;3}. For each models, the deviance was generated. In a first step we compared all first-order polynomials and then all second-order to the linear model (model with a first-order fractional polynomial with the power p=1).

For the number of cigarettes per day the best model was defined with powers p1=-2 and p2=2, $FP_2 = \beta_0 + \beta_1 (\text{Num\_Cigarettes})^{-2} + \beta_2 (\text{Num\_Cigarettes})^{2}$. And the same powers was found for the duration of smoking $FP_2 = \beta_0 + \beta_1 (\text{Years of smoking})^{-2} + \beta_2 (\text{Years of smoking})^{2}$

In the second stage, the pooled estimation was calculated using a random effects model and the moment estimation method.

For categorical variables, heterogeneity between studies was evaluated using the Q test statistic.

For categorical variables, we tested the linear trends across levels of cigarette smoking; we first estimated trends in each study and used the Wald test to estimate the P value of the summary variable from the random-effects models (ref Smith-Warner 2006). To investigate whether the effect of cigarette smoking was homogenous across strata of selected covariates, we conducted analyses stratified by age, sex and geographic area. Heterogeneity across strata was assessed using the Q test statistic.

Pooled estimations were generated using R software and the function metagen from the library “meta” (See supplements).

Corresponding graphics and forest plot were created using R software and the library “gplot”.

We also conducted a sensitivity analysis to evaluate the influence of *Helicobacter pylori (HP)* infection information by excluding all studies without the information in a first time and in a second time considering only *HP* positive controls.

4. Results

4.1. Category of smoker

The pooled ORs for gastric cancer according to cigarette smoking habits are given in Table 4. Concerning studies where the former status was available, ORs was 1.19 (95% CI 1.08-1.30) for ever cigarette smokers, 1.14 (95% CI 1.01-1.29) for former cigarette smokers and 1.22 (95% CI 1.06-1.40) for current smokers, compared with never smokers. Among current smokers, the risk increased with categories of the number of cigarettes smoked per day. Compared to never smokers, ORs were 1.05 (95% CI 0.88-1.26) for 0 to 10 cigarettes per day, 1.27 (95% CI 1.11-1.45) for 10 to 20 cigarettes per day and 1.29 (95% CI 1.06-1.57) for more than 20 cigarettes per day, with a significant trend (p=0.005). The risk increased also significantly with increasing duration of smoking (*p value* for trend p<0.0001) and with ORs in category equal to 1.04 (95%CI 0.94-1.16) for less than 30 years of smoking, 1.32 (95% CI 1.16-1.49) for a duration between 30 to 40 years of smoking and 1.33 (95% CI 1.14-1.54) for more than 30 years of cigarette smoking. A significant decreasing trend in risk was found with an increase time since stopping cigarette smoking (p=0.02) taking as reference current smokers (Table 4).

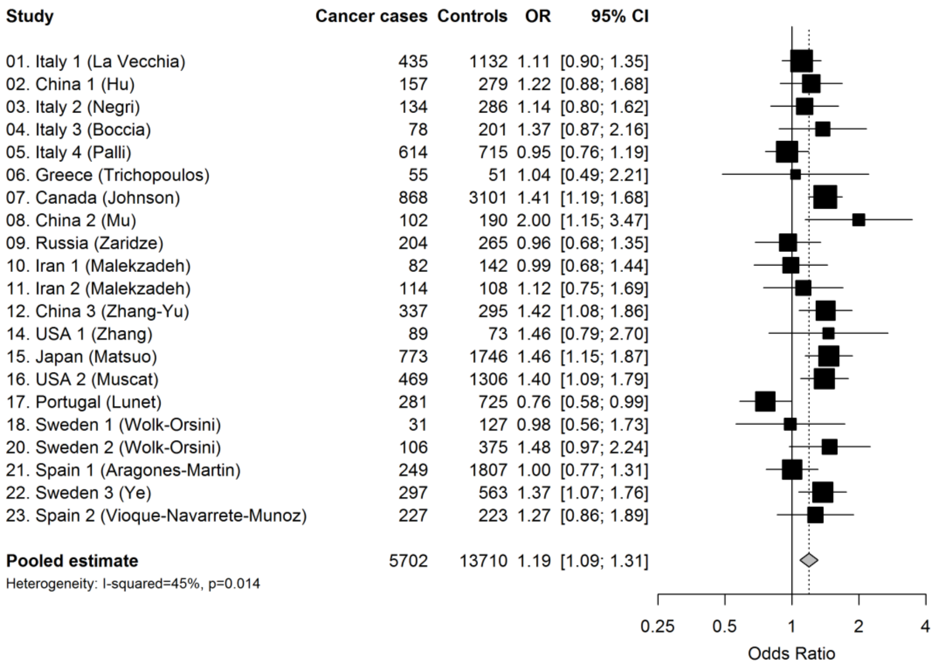
Table 4. Pooled odds ratios (ORs) and 95% confidence intervals (CIs) for gastric cancer according to cigarette and tobacco smoking habits. Stomach cancer pooling (StoP) consortium.

	Cases		**Controls**		**OR[1] (CI 95%)**
	N	**%**	**N**	**%**	
Total	10039		25596		
Cigarette smoking status					
Never smoker	4122	41.1	11396	44.5	1
Ever cigarette smoker	5510	54.8	13516	52.8	1.19 (1.08-1.30)
Former cigarette smoker	2775	27.6	7421	29.0	1.14 (1.01-1.29)
Current cigarette smoker	2735	27.2	6095	23.8	1.22 (1.06-1.40)
Other than cigarette smoker	121	1.2	343	1.3	1.09 (0.79-1.50)
Missing	*288*	*2.9*	*350*	*1.4*	
Intensity (cigarettes per day)[3]					
0 to ≤10	674	6.7	1820	7.1	1.05 (0.88-1.26)
>10 to ≤20	1285	12.8	2696	10.5	1.27 (1.11-1.45)
> 20	748	7.5	1497	5.8	1.29 (1.06-1.57)
missing	*316*	*3.2*	*432*	*1.7*	
P value for trend					0.005
Cigarette smoking duration (years)					
0 to ≤30	2213	22.0	6921	27.0	1.04 (0.94-1.16)
>30 to ≤40	1420	14.1	3031	11.8	1.32 (1.16-1.49)
> 40	1661	16.5	3009	11.8	1.33 (1.14-1.54)
missing	*504*	*5.0*	*905*	*3.5*	
P value for trend					<0.0001
Total[3]	7657		18222		
Time since stopping cigarette smoking (years)					
Never smoker	3204	41.8	8212	45.1	1
0 to <10	674	8.8	1543	8.5	1.15 (0.95-1.39)
10 to <20	513	6.7	1391	7.6	1.07 (0.94-1.23)
≥ 20	616	8.0	1718	9.4	1.03 (0.87-1.21)
Other than cigarette smoker	121	1.6	343	1.9	
Missing	*280*	*3.7*	*482*	*2.6*	
P value for trend					0.1628
Time since stopping cigarette smoking (years)					
Current cigarette smoker	2249	29.4	4533	24.9	1
0 to <10	674	8.8	1543	8.5	0.92 (0.73-1.16)
10 to <20	513	6.7	1391	7.6	0.82 (0.72-0.94)
≥ 20	616	8.0	1718	9.4	0.84 (0.66-1.07)
Other than cigarette smoker	121	1.6	343	1.9	

Missing	*280*	*3.7*	*482*	*2.6*	
P value for trend					0.018

[1]Pooled ORs were computed using random-effects models, study-specific ORs were adjusted, when available, for sex, age, race/ethnicity, social class, alcohol drinking, fruit and vegetable consumption and study center for multicentric studies. [2]Cigarette smoking status was not available studies China 4 (Zhang) and Iran 3 (Malekzadeh). [3]Current smokers only [4]Time since stopping cigarette smoking was not available for studies Greece (Trichopoulos), Canada (Johnson), China 1 (Mu), Iran 1 (Malekzadeh), Iran 2 (Malekzadeh), USA 1 (Zhang), Sweden 1 (Wolk-Orsini), and Sweden 2 (Wolk-Orsini)

A forest plot of the study-specific and the pooled ORs for gastric cancer risk for ever smokers compared to never smokers are given in Figure 5.



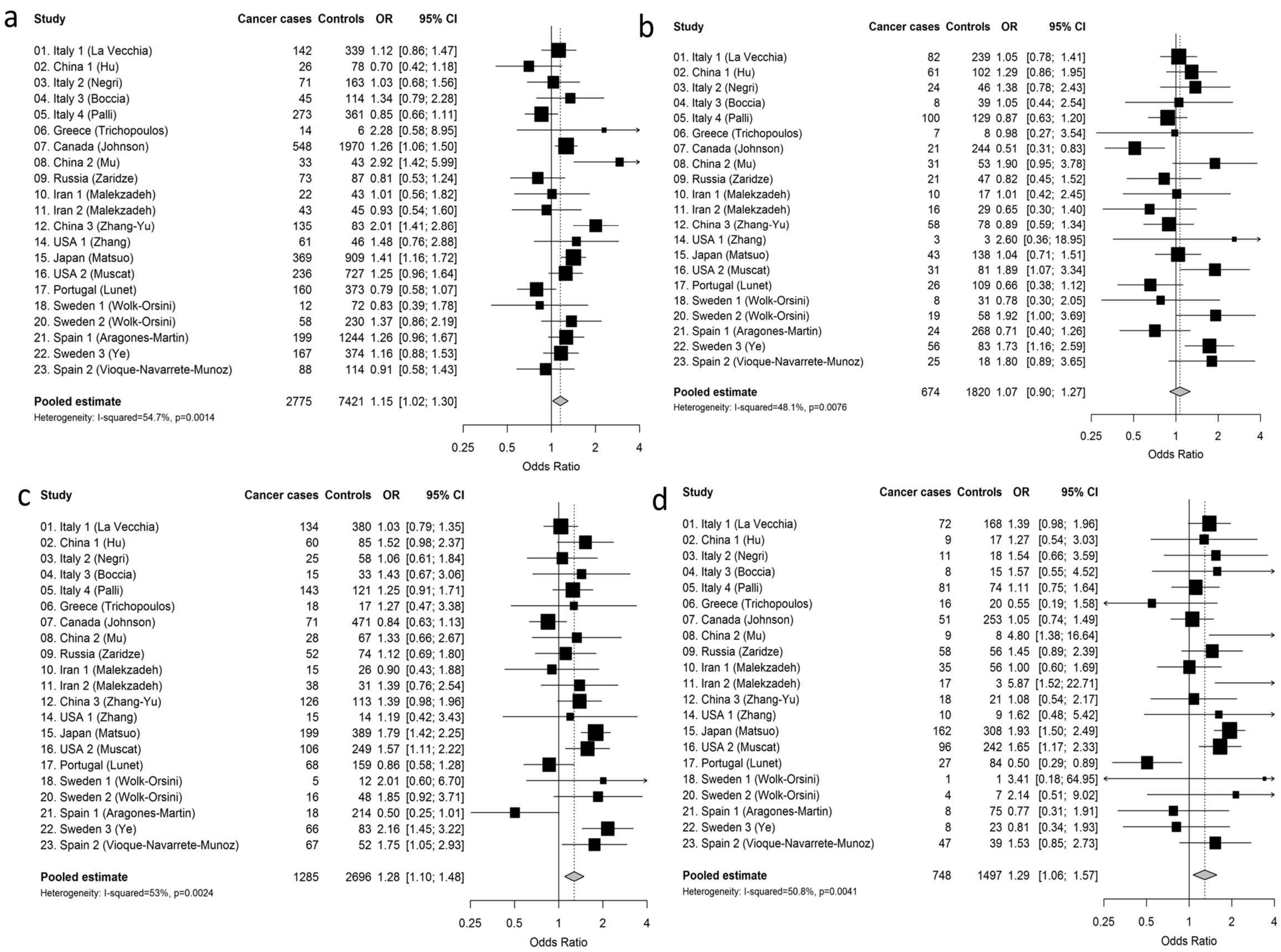
Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	435	1132	1.11	[0.90; 1.35]
02. China 1 (Hu)	157	279	1.22	[0.88; 1.68]
03. Italy 2 (Negri)	134	286	1.14	[0.80; 1.62]
04. Italy 3 (Boccia)	78	201	1.37	[0.87; 2.16]
05. Italy 4 (Palli)	614	715	0.95	[0.76; 1.19]
06. Greece (Trichopoulos)	55	51	1.04	[0.49; 2.21]
07. Canada (Johnson)	868	3101	1.41	[1.19; 1.68]
08. China 2 (Mu)	102	190	2.00	[1.15; 3.47]
09. Russia (Zaridze)	204	265	0.96	[0.68; 1.35]
10. Iran 1 (Malekzadeh)	82	142	0.99	[0.68; 1.44]
11. Iran 2 (Malekzadeh)	114	108	1.12	[0.75; 1.69]
12. China 3 (Zhang-Yu)	337	295	1.42	[1.08; 1.86]
14. USA 1 (Zhang)	89	73	1.46	[0.79; 2.70]
15. Japan (Matsuo)	773	1746	1.46	[1.15; 1.87]
16. USA 2 (Muscat)	469	1306	1.40	[1.09; 1.79]
17. Portugal (Lunet)	281	725	0.76	[0.58; 0.99]
18. Sweden 1 (Wolk-Orsini)	31	127	0.98	[0.56; 1.73]
20. Sweden 2 (Wolk-Orsini)	106	375	1.48	[0.97; 2.24]
21. Spain 1 (Aragones-Martin)	249	1807	1.00	[0.77; 1.31]
22. Sweden 3 (Ye)	297	563	1.37	[1.07; 1.76]
23. Spain 2 (Vioque-Navarrete-Munoz)	227	223	1.27	[0.86; 1.89]
Pooled estimate	5702	13710	1.19	[1.09; 1.31]

Heterogeneity: I-squared=45%, p=0.014

Figure 5: Pooled OR and corresponding 95% confidence interval for gastric cancer risk for ever smokers compared to never smokers, Stomach cancer pooling (StoP) consortium.

4.2. Number of cigarettes per day

A forest plot of the study-specific and the pooled ORs for gastric cancer risk for numbers of cigarettes smoked per day among current smoker compared to never smokers are given in Figure 6. The heterogeneity across study was significant for each category of consumption.



a

Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	142	339	1.12	[0.86; 1.47]
02. China 1 (Hu)	26	78	0.70	[0.42; 1.18]
03. Italy 2 (Negri)	71	163	1.03	[0.68; 1.56]
04. Italy 3 (Boccia)	45	114	1.34	[0.79; 2.28]
05. Italy 4 (Palli)	273	361	0.85	[0.66; 1.11]
06. Greece (Trichopoulos)	14	6	2.28	[0.58; 8.95]
07. Canada (Johnson)	548	1970	1.26	[1.06; 1.50]
08. China 2 (Mu)	33	43	2.92	[1.42; 5.99]
09. Russia (Zaridze)	73	87	0.81	[0.53; 1.24]
10. Iran 1 (Malekzadeh)	22	43	1.01	[0.56; 1.82]
11. Iran 2 (Malekzadeh)	43	45	0.93	[0.54; 1.60]
12. China 3 (Zhang-Yu)	135	83	2.01	[1.41; 2.86]
14. USA 1 (Zhang)	61	46	1.48	[0.76; 2.88]
15. Japan (Matsuo)	369	909	1.41	[1.16; 1.72]
16. USA 2 (Muscat)	236	727	1.25	[0.96; 1.64]
17. Portugal (Lunet)	160	373	0.79	[0.58; 1.07]
18. Sweden 1 (Wolk-Orsini)	12	72	0.83	[0.39; 1.78]
20. Sweden 2 (Wolk-Orsini)	58	230	1.37	[0.86; 2.19]
21. Spain 1 (Aragones-Martin)	199	1244	1.26	[0.96; 1.67]
22. Sweden 3 (Ye)	167	374	1.16	[0.88; 1.53]
23. Spain 2 (Vioque-Navarrete-Munoz)	88	114	0.91	[0.58; 1.43]
Pooled estimate	2775	7421	1.15	[1.02; 1.30]

Heterogeneity: I-squared=54.7%, p=0.0014

b

Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	82	239	1.05	[0.78; 1.41]
02. China 1 (Hu)	61	102	1.29	[0.86; 1.95]
03. Italy 2 (Negri)	24	46	1.38	[0.78; 2.43]
04. Italy 3 (Boccia)	8	39	1.05	[0.44; 2.54]
05. Italy 4 (Palli)	100	129	0.87	[0.63; 1.20]
06. Greece (Trichopoulos)	7	8	0.98	[0.27; 3.54]
07. Canada (Johnson)	21	244	0.51	[0.31; 0.83]
08. China 2 (Mu)	31	53	1.90	[0.95; 3.78]
09. Russia (Zaridze)	21	47	0.82	[0.45; 1.52]
10. Iran 1 (Malekzadeh)	10	17	1.01	[0.42; 2.45]
11. Iran 2 (Malekzadeh)	16	29	0.65	[0.30; 1.40]
12. China 3 (Zhang-Yu)	58	78	0.89	[0.59; 1.34]
14. USA 1 (Zhang)	3	3	2.60	[0.36; 18.95]
15. Japan (Matsuo)	43	138	1.04	[0.71; 1.51]
16. USA 2 (Muscat)	31	81	1.89	[1.07; 3.34]
17. Portugal (Lunet)	26	109	0.66	[0.38; 1.12]
18. Sweden 1 (Wolk-Orsini)	8	31	0.78	[0.30; 2.05]
20. Sweden 2 (Wolk-Orsini)	19	58	1.92	[1.00; 3.69]
21. Spain 1 (Aragones-Martin)	24	268	0.71	[0.40; 1.26]
22. Sweden 3 (Ye)	56	83	1.73	[1.16; 2.59]
23. Spain 2 (Vioque-Navarrete-Munoz)	25	18	1.80	[0.89; 3.65]
Pooled estimate	674	1820	1.07	[0.90; 1.27]

Heterogeneity: I-squared=48.1%, p=0.0076

c

Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	134	380	1.03	[0.79; 1.35]
02. China 1 (Hu)	60	85	1.52	[0.98; 2.37]
03. Italy 2 (Negri)	25	58	1.06	[0.61; 1.84]
04. Italy 3 (Boccia)	15	33	1.43	[0.67; 3.06]
05. Italy 4 (Palli)	143	121	1.25	[0.91; 1.71]
06. Greece (Trichopoulos)	18	17	1.27	[0.47; 3.38]
07. Canada (Johnson)	71	471	0.84	[0.63; 1.13]
08. China 2 (Mu)	28	67	1.33	[0.66; 2.67]
09. Russia (Zaridze)	52	74	1.12	[0.69; 1.80]
10. Iran 1 (Malekzadeh)	15	26	0.90	[0.43; 1.88]
11. Iran 2 (Malekzadeh)	38	31	1.39	[0.76; 2.54]
12. China 3 (Zhang-Yu)	126	113	1.39	[0.98; 1.96]
14. USA 1 (Zhang)	15	14	1.19	[0.42; 3.43]
15. Japan (Matsuo)	199	389	1.79	[1.42; 2.25]
16. USA 2 (Muscat)	106	249	1.57	[1.11; 2.22]
17. Portugal (Lunet)	68	159	0.86	[0.58; 1.28]
18. Sweden 1 (Wolk-Orsini)	5	12	2.01	[0.60; 6.70]
20. Sweden 2 (Wolk-Orsini)	16	48	1.85	[0.92; 3.71]
21. Spain 1 (Aragones-Martin)	18	214	0.50	[0.25; 1.01]
22. Sweden 3 (Ye)	66	83	2.16	[1.45; 3.22]
23. Spain 2 (Vioque-Navarrete-Munoz)	67	52	1.75	[1.05; 2.93]
Pooled estimate	1285	2696	1.28	[1.10; 1.48]

Heterogeneity: I-squared=53%, p=0.0024

d

Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	72	168	1.39	[0.98; 1.96]
02. China 1 (Hu)	9	17	1.27	[0.54; 3.03]
03. Italy 2 (Negri)	11	18	1.54	[0.66; 3.59]
04. Italy 3 (Boccia)	8	15	1.57	[0.55; 4.52]
05. Italy 4 (Palli)	81	74	1.11	[0.75; 1.64]
06. Greece (Trichopoulos)	16	20	0.55	[0.19; 1.58]
07. Canada (Johnson)	51	253	1.05	[0.74; 1.49]
08. China 2 (Mu)	9	8	4.80	[1.38; 16.64]
09. Russia (Zaridze)	58	56	1.45	[0.89; 2.39]
10. Iran 1 (Malekzadeh)	35	56	1.00	[0.60; 1.69]
11. Iran 2 (Malekzadeh)	17	3	5.87	[1.52; 22.71]
12. China 3 (Zhang-Yu)	18	21	1.08	[0.54; 2.17]
14. USA 1 (Zhang)	10	9	1.62	[0.48; 5.42]
15. Japan (Matsuo)	162	308	1.93	[1.50; 2.49]
16. USA 2 (Muscat)	96	242	1.65	[1.17; 2.33]
17. Portugal (Lunet)	27	84	0.50	[0.29; 0.89]
18. Sweden 1 (Wolk-Orsini)	1	1	3.41	[0.18; 64.95]
20. Sweden 2 (Wolk-Orsini)	4	7	2.14	[0.51; 9.02]
21. Spain 1 (Aragones-Martin)	8	75	0.77	[0.31; 1.91]
22. Sweden 3 (Ye)	8	23	0.81	[0.34; 1.93]
23. Spain 2 (Vioque-Navarrete-Munoz)	47	39	1.53	[0.85; 2.73]
Pooled estimate	748	1497	1.29	[1.06; 1.57]

Heterogeneity: I-squared=50.8%, p=0.0041

Figure 6: Pooled OR and corresponding 95% confidence interval for gastric cancer risk for former smokers (a), smokers of less than 10 cigarettes (b), smokers of 10 to 20 cigarettes (c) and smokers of more than 20 cigarettes (d) compared to never smokers, Stomach cancer pooling (StoP) consortium.

Figure 7 represented the relation between gastric cancer risk and the number of cigarette smoking. The relation is fitted by a fractional polynomial. For the number of cigarettes per day the best fitting model was defined with powers p1=-2 and p2=2, $FP_2= \beta_0 + \beta_1 X^{-2} + \beta_2 X^2$. This curve represented the increasing risk of gastric cancer with increasing risk of smoking cigarettes per day. It showed that the risk increased slightly for up to 2 packs of cigarettes and the increase appeared to be stronger after. The fractional polynomial is significantly different from linear model ($p<0.0001$)

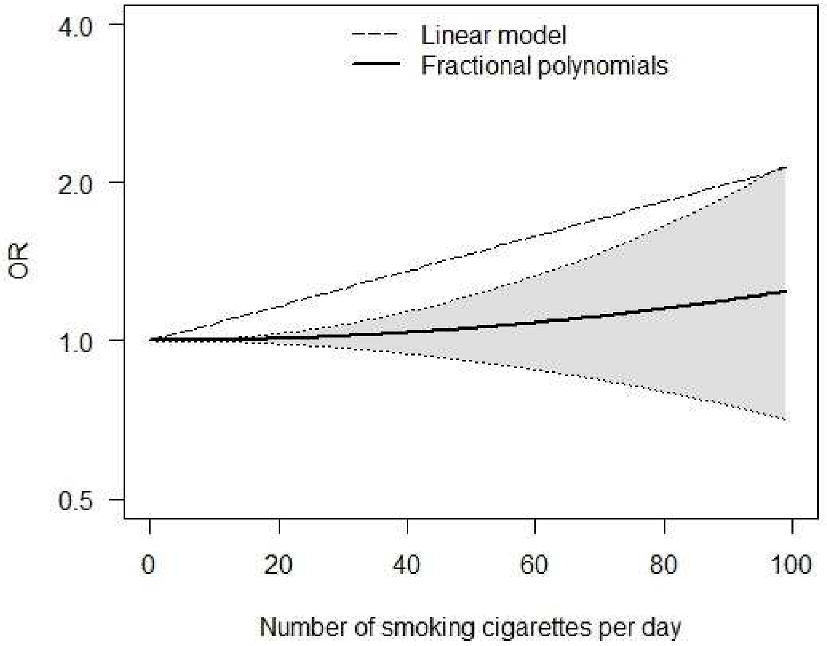


Figure 7: Relation between number of smoking cigarettes per day and risk of gastric cancer fitted by a fractional polynomial $FP_2= \beta_0 + \beta_1 X^{-2} + \beta_2 X^2$ and a linear model, Stomach cancer pooling (StoP) consortium.

Table 5 showed a contrast between results across categorical model, linear model and fractional polynomials.

Table 5. Contrast of pooled odds ratios (ORs) and 95% confidence intervals (CIs) for gastric cancer estimated according to the number of smoking cigarettes per day and in continuous through a linear model and second-order fractional polynomials. Stomach cancer pooling (StoP) consortium.

Cigarettes per day					
Categorization		**Fractional polynomials**			
Range	**OR cat**	**Range**	**Ref. point**	**OR linear**	**OR FP2**
Never	1.00	Never	0	1.00	1.00
1 – 10	1.07 (0.90-1.27)	1 – 10	5	1.03 (1.02-1.04)	1.00 (0.99-1.00)
11 to 20	1.28 (1.10-1.48)	11 – 30	15	1.11 (1.07-1.14)	1.01 (0.99-1.02)
21 +	1.29 (1.06-1.57)	31 – 50	25	1.19 (1.11-1.25)	1.01 (0.98-1.05)
		51 – 60	35	1.27 (1.17-1.37)	1.03 (0.96-1.10)
		61 – 70	45	1.36 (1.22-1.50)	1.05 (0.93-1.18)
		71 – 80	55	1.46 (1.28-1.65)	1.07 (0.89-1.28)
			65	1.56 (1.33-1.81)	1.10 (0.85-1.42)
			75	1.67 (1.40-1.98)	1.14 (0.81-1.59)

The association between the number of cigarettes smoked in category and gastric cancer risk was further assessed in strata of sex, age and geographic area. We noticed a stronger effect of duration for men in for smokers for less than 10 years, for young smokers for more than 30 years (Figure 8). Similar risks were found for cardia and non-cardia gastric cancer cases. Considering only studies with the information on *HP* infection, effects of cigarette smoking did not materially when taking into account controls with a positive *HP* infection test in our analyses. We further considerate separately controls recruited from hospital and those recruited in the general population. Risks appeared to be slightly higher in the analyses with hospital controls.

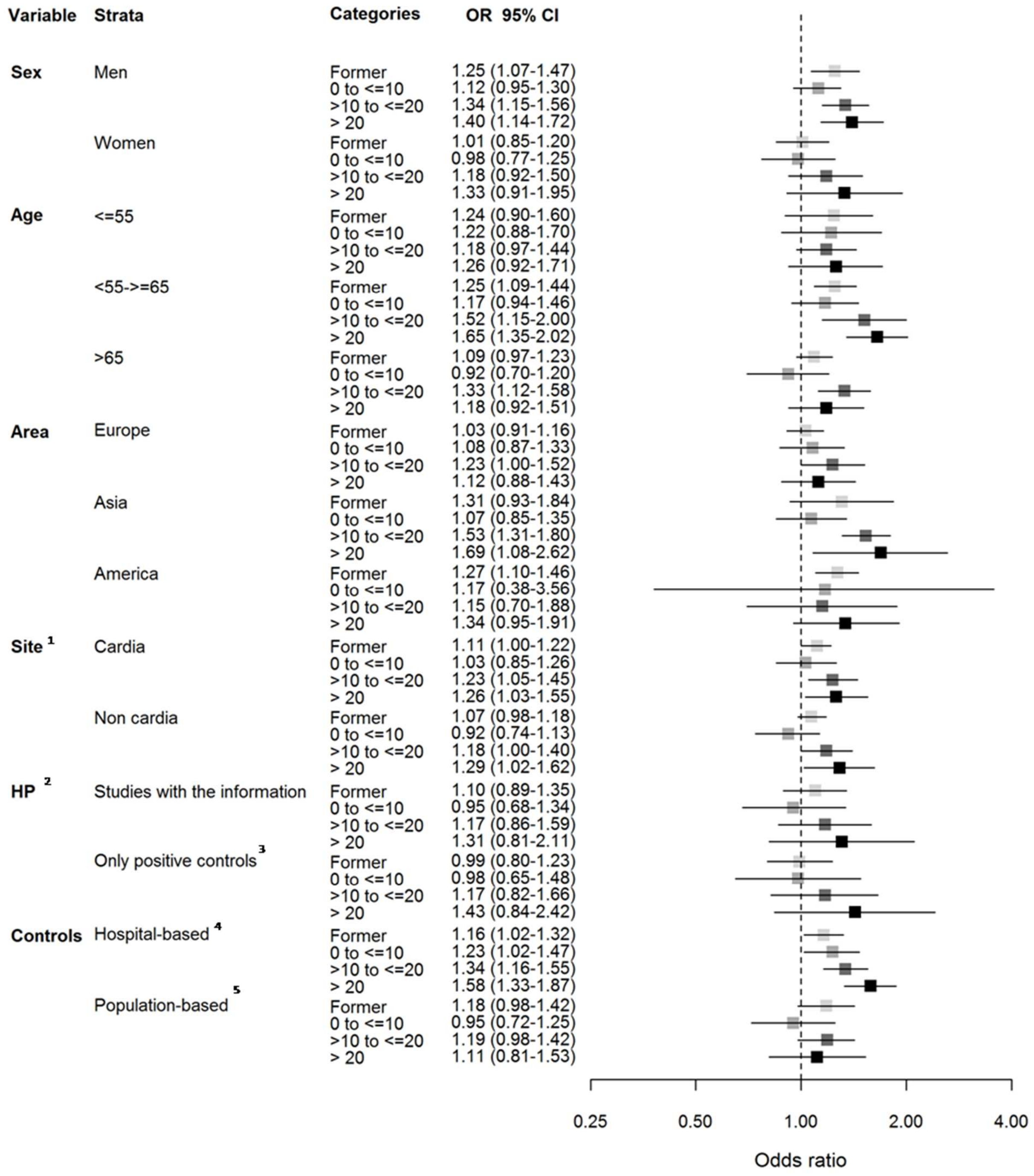


Figure 8: Pooled odds ratios (ORs)1 and 95% confidence intervals (CIs) for gastric cancer according to cigarette smoking status in strata of sex, age, geographic area, cancer site, Helicobacter Pylori infection, controls recruitment. Stomach cancer pooling (StoP) consortium.

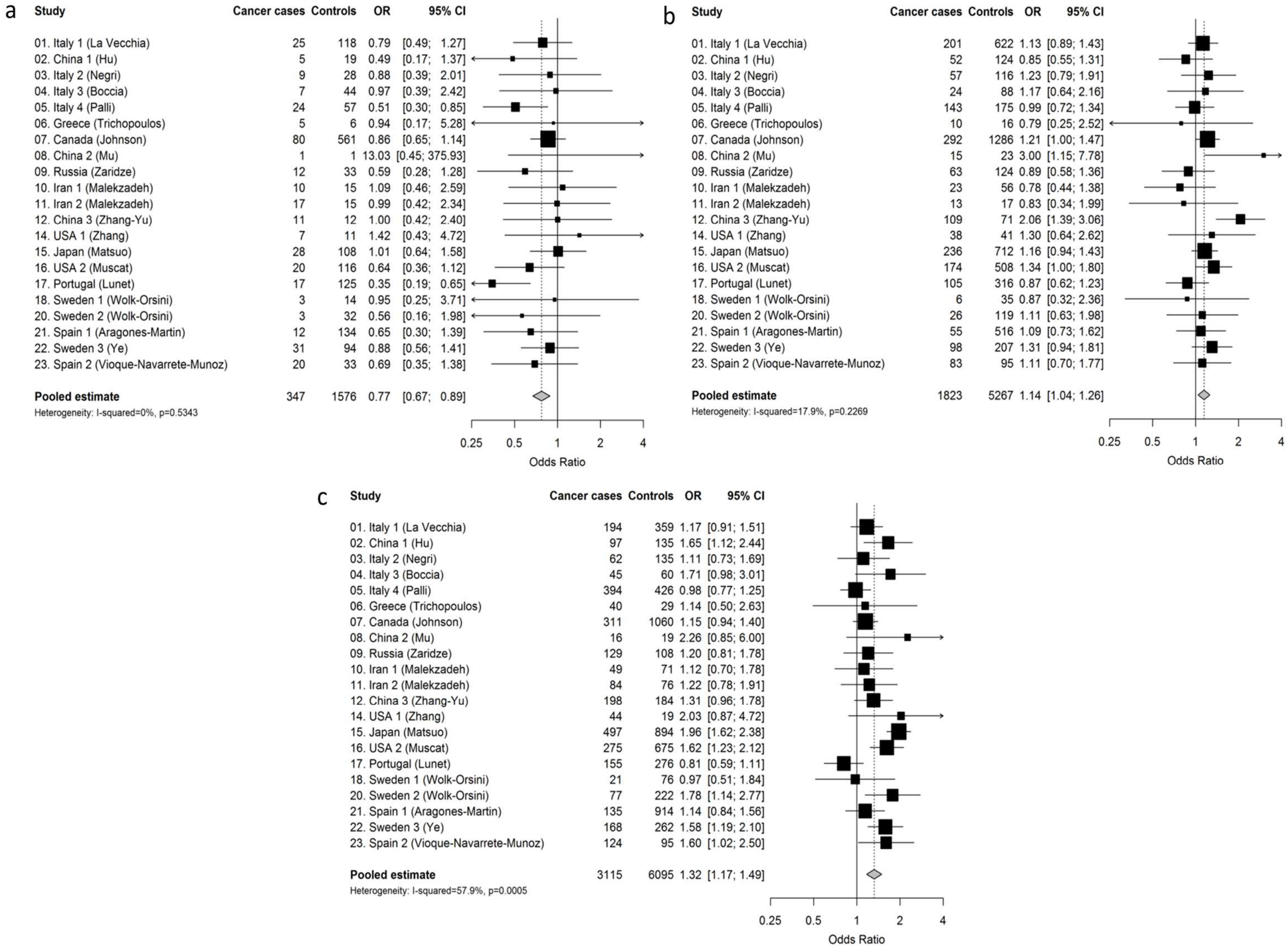
[1] The study Italy 3 (Boccia) and Spain 2 (Vioque) were not considered because controls were all *HP* negative
Considered studies : China 2 (Mu), Iran 1, Iran 2 (Malekzadeh), Japan (Matsuo), Portugal (Lunet), Russia (Zaridze), Spain 1 (Aragones-Martin), Sweden 3 (Ye)

[2] Considered studies: Italy 1 (La Vecchia), Italy 2 (Negri), Italy 3 (Boccia), Italy 4 (Palli), Canada (Johnson), Russia (Zaridze), Iran 1, Iran 2 (Malekzadeh), USA 1 (Zhang), Japan (Matsuo), USA 2 (Muscat), Portugal (lunet), Sweden 1, Sweden 2 (Wolk-Orsini), Spain 1 (Aragones-Martin), Sweden 3 (Ye), Spain 2 (Vioque)

The study Greece (Trichopoulos) was not considered because all its cases had a non-cardia neoplasm
For the category 10-20 cig/day, the study 14 USA 1 (Zhang) was not considered because of lack of cases
For the category >20 cig/day, studies 01.Italy 1 (La Vecchia) and 18.Sweden 1 were not considered because of lack of cases.
[3] Pooled ORs were computed considering only controls with a positive test of *helicobacter pylori* infection
[4] Considered studies: Italy 1 (La Vecchia), China 1 (Hu), Italy 2 (Negri), Italy 3 (Boccia), Greece (Trichopoulos), USA 1 (Zhang), Japan (Matsuo), USA 2 (Muscat), Spain 2 (Vioque)
[5] Considered studies: Italy 4 (Palli), Canada (Johnson), China 2 (Mu), Iran 1, Iran 2 (Malekzadeh), China 3 (Zhang-Yu), Portugal (lunet), Sweden 1, Sweden 2 (Wolk-Orsini), Spain 1 (Aragones-Martin), Sweden 3 (Ye)
The study Russia (Zaridze) was not considered in this analysis because it considers both hospital and general population controls

4.3. Duration of smoking in years.

A forest plot of the study-specific and the pooled ORs for gastric cancer risk for the duration of smoking cigarettes compared to never smokers are given in Figure 9. The heterogeneity across study was significant only for the category of smokers for more than 30 years.



a

Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	25	118	0.79	[0.49; 1.27]
02. China 1 (Hu)	5	19	0.49	[0.17; 1.37]
03. Italy 2 (Negri)	9	28	0.88	[0.39; 2.01]
04. Italy 3 (Boccia)	7	44	0.97	[0.39; 2.42]
05. Italy 4 (Palli)	24	57	0.51	[0.30; 0.85]
06. Greece (Trichopoulos)	5	6	0.94	[0.17; 5.28]
07. Canada (Johnson)	80	561	0.86	[0.65; 1.14]
08. China 2 (Mu)	1	1	13.03	[0.45; 375.93]
09. Russia (Zaridze)	12	33	0.59	[0.28; 1.28]
10. Iran 1 (Malekzadeh)	10	15	1.09	[0.46; 2.59]
11. Iran 2 (Malekzadeh)	17	15	0.99	[0.42; 2.34]
12. China 3 (Zhang-Yu)	11	12	1.00	[0.42; 2.40]
14. USA 1 (Zhang)	7	11	1.42	[0.43; 4.72]
15. Japan (Matsuo)	28	108	1.01	[0.64; 1.58]
16. USA 2 (Muscat)	20	116	0.64	[0.36; 1.12]
17. Portugal (Lunet)	17	125	0.35	[0.19; 0.65]
18. Sweden 1 (Wolk-Orsini)	3	14	0.95	[0.25; 3.71]
20. Sweden 2 (Wolk-Orsini)	3	32	0.56	[0.16; 1.98]
21. Spain 1 (Aragones-Martin)	12	134	0.65	[0.30; 1.39]
22. Sweden 3 (Ye)	31	94	0.88	[0.56; 1.41]
23. Spain 2 (Vioque-Navarrete-Munoz)	20	33	0.69	[0.35; 1.38]
Pooled estimate	347	1576	0.77	[0.67; 0.89]

Heterogeneity: I-squared=0%, p=0.5343

b

Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	201	622	1.13	[0.89; 1.43]
02. China 1 (Hu)	52	124	0.85	[0.55; 1.31]
03. Italy 2 (Negri)	57	116	1.23	[0.79; 1.91]
04. Italy 3 (Boccia)	24	88	1.17	[0.64; 2.16]
05. Italy 4 (Palli)	143	175	0.99	[0.72; 1.34]
06. Greece (Trichopoulos)	10	16	0.79	[0.25; 2.52]
07. Canada (Johnson)	292	1286	1.21	[1.00; 1.47]
08. China 2 (Mu)	15	23	3.00	[1.15; 7.78]
09. Russia (Zaridze)	63	124	0.89	[0.58; 1.36]
10. Iran 1 (Malekzadeh)	23	56	0.78	[0.44; 1.38]
11. Iran 2 (Malekzadeh)	13	17	0.83	[0.34; 1.99]
12. China 3 (Zhang-Yu)	109	71	2.06	[1.39; 3.06]
14. USA 1 (Zhang)	38	41	1.30	[0.64; 2.62]
15. Japan (Matsuo)	236	712	1.16	[0.94; 1.43]
16. USA 2 (Muscat)	174	508	1.34	[1.00; 1.80]
17. Portugal (Lunet)	105	316	0.87	[0.62; 1.23]
18. Sweden 1 (Wolk-Orsini)	6	35	0.87	[0.32; 2.36]
20. Sweden 2 (Wolk-Orsini)	26	119	1.11	[0.63; 1.98]
21. Spain 1 (Aragones-Martin)	55	516	1.09	[0.73; 1.62]
22. Sweden 3 (Ye)	98	207	1.31	[0.94; 1.81]
23. Spain 2 (Vioque-Navarrete-Munoz)	83	95	1.11	[0.70; 1.77]
Pooled estimate	1823	5267	1.14	[1.04; 1.26]

Heterogeneity: I-squared=17.9%, p=0.2269

c

Study	Cancer cases	Controls	OR	95% CI
01. Italy 1 (La Vecchia)	194	359	1.17	[0.91; 1.51]
02. China 1 (Hu)	97	135	1.65	[1.12; 2.44]
03. Italy 2 (Negri)	62	135	1.11	[0.73; 1.69]
04. Italy 3 (Boccia)	45	60	1.71	[0.98; 3.01]
05. Italy 4 (Palli)	394	426	0.98	[0.77; 1.25]
06. Greece (Trichopoulos)	40	29	1.14	[0.50; 2.63]
07. Canada (Johnson)	311	1060	1.15	[0.94; 1.40]
08. China 2 (Mu)	16	19	2.26	[0.85; 6.00]
09. Russia (Zaridze)	129	108	1.20	[0.81; 1.78]
10. Iran 1 (Malekzadeh)	49	71	1.12	[0.70; 1.78]
11. Iran 2 (Malekzadeh)	84	76	1.22	[0.78; 1.91]
12. China 3 (Zhang-Yu)	198	184	1.31	[0.96; 1.78]
14. USA 1 (Zhang)	44	19	2.03	[0.87; 4.72]
15. Japan (Matsuo)	497	894	1.96	[1.62; 2.38]
16. USA 2 (Muscat)	275	675	1.62	[1.23; 2.12]
17. Portugal (Lunet)	155	276	0.81	[0.59; 1.11]
18. Sweden 1 (Wolk-Orsini)	21	76	0.97	[0.51; 1.84]
20. Sweden 2 (Wolk-Orsini)	77	222	1.78	[1.14; 2.77]
21. Spain 1 (Aragones-Martin)	135	914	1.14	[0.84; 1.56]
22. Sweden 3 (Ye)	168	262	1.58	[1.19; 2.10]
23. Spain 2 (Vioque-Navarrete-Munoz)	124	95	1.60	[1.02; 2.50]
Pooled estimate	3115	6095	1.32	[1.17; 1.49]

Heterogeneity: I-squared=57.9%, p=0.0005

Figure 9: Pooled OR and corresponding 95% confidence interval for gastric cancer risk for subjects smoking less than 10 years (a), between 10 and 30 years (b) and more than 30 years (c) compared to never smokers, Stomach cancer pooling (StoP) consortium.

Figure 10 represented the increasing risk of gastric cancer with increasing duration of smoking fitted by a linear model and by a fractional polynomial. The best fitting model was defined with powers p1=-2 and p2=2, $FP_2 = \beta_0 + \beta_1 X^{-2} + \beta_2 X^2$. This graph provided the evidence of strong non linear dose relationship between risk of gastric cancer and increasing duration of cigarette smoking. The fractional polynomial is significantly different from linear model ($p<0.0001$) AIC=-645.5. The fractional polynomial suggested a stronger increase of risk after 20 years of smoking.

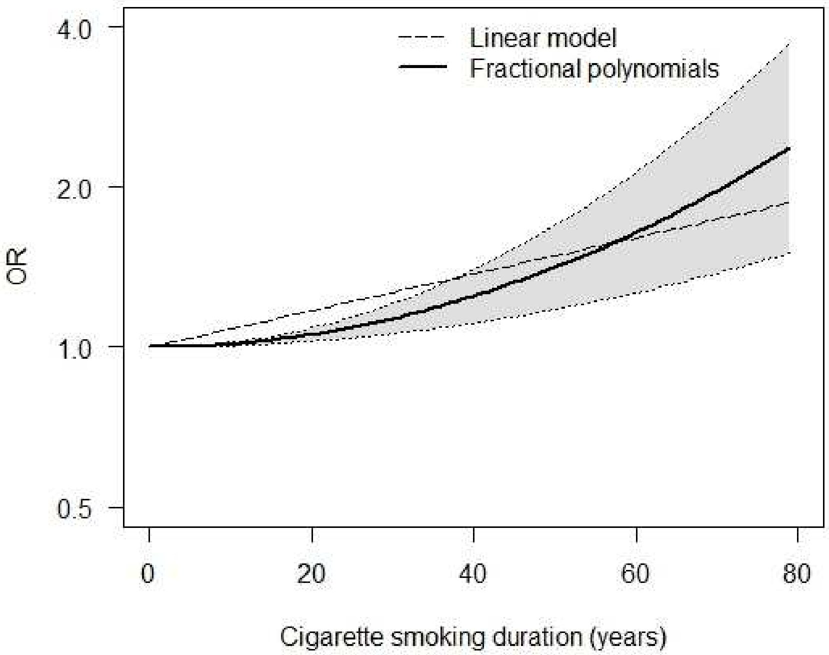


Figure 10: Relation between duration of smoking and risk of gastric cancer fitted by a fractional polynomial $FP_2 = \beta_0 + \beta_1 X^{-2} + \beta_2 X^2$, and a linear model, Stomach cancer pooling (StoP) consortium.

Table 6 showed a contrast between results across categorical model, linear model and fractional polynomials.

Table 6. Contrast of pooled odds ratios (ORs) and 95% confidence intervals (CIs) for gastric cancer estimated according to the number of smoking cigarettes per day and in continuous through a linear model and second-order fractional polynomials. Stomach cancer pooling (StoP) consortium.

Cigarette smoking duration (years)					
Categorisation		**Fractional polynomials**			
Range	**OR cat**	**Range**	**Ref. point**	**OR linear**	**OR FP2**
Never		Never	0	1.00	1.00
1 – 10	0.77 (0.67-0.89)	1 – 10	5	1.04 (1.03-1.04)	1.00 (1.00-1.00)
11 to 30	1.14 (1.04-1.26)	11 – 30	20	1.17 (1.13-1.20)	1.06 (1.02-1.08)
31 +	1.32 (1.17-1.49)	31 – 50	40	1.37 (1.28-1.45)	1.25 (1.10- 1.38)
		51 – 60	55	1.54 (1.40-1.67)	1.52 (1.21-1.85)
		61 – 70	65	1.67 (1.49-1.84)	1.79 (1.31-1.85)
		71 – 80	75	1.81 (1.59-2.02)	2.17 (1.43-3.17)

The association between smoking duration and gastric cancer risk was further assessed in strata of sex, age and geographic area. We noticed a stronger effect of duration for men in for smokers for less than 10 years, for young smokers for more than 30 years (Table 7).

Table 7. Pooled odds ratios (ORs) and 95% confidence intervals (CIs) for gastric cancer according to cigarette smoking duration in strata of selected covariates among 10040 cases and 25602 controls. Stomach cancer pooling (StoP) consortium.

	Never	**0-≤10 years**		**>10-≤30 years**		**>30 years**	
	Ca : Co	**Ca : Co**	**OR (95% CI)**	**Ca : Co**	**OR (95% CI)**	**Ca : Co**	**OR (95% CI)**
Overall	4120:11390	347:1576	0.77 (0.67-0.89)	1823:5267	1.14 (1.04-1.26)	3115:6095	1.32 (1.17-1.49)
Sex							
Men	1597:4484	270:978	0.86 (0.73-1.02)	1482: 3824	1.21 (1.06-1.38)	2751:4942	1.39 (1.20-1.60)
Women	2523:6906	77:598	0.58 (0.43-0.77)	341:1443	1.11 (0.94-1.30)	364:1153	1.17 (0.98-1.40)
P for interaction			0.0183		0.3977		0.1482
Age							
≤55	961:3624	124:826	0.74 (0.58-0.95)	842:2678	1.20 (0.98-1.48)	427:982	1.50 (1.19-1.90)
<55 - ≤65	1076:3152	102:358	0.75 (0.59-0.96)	486:1326	1.20 (1.02-1.41)	1148:2224	1.46 (1.24-1.71)
>65	2083:4610	121:392	0.99 (0.74-1.33)	4951263	1.03 (0.88-1.19)	15402889	1.15 (1.02-1.30)
P for interaction			0.2587		0.2884		0.0295
Geographic area							
America	551:2644	107:688	0.83 (0.65-1.06)	504:1835	1.25 (1.07-1.47)	630:1754	1.40 (1.04-1.91)
Asia	1358:2884	72:170	1.00 (0.71-1.42)	448:1003	1.20 (0.82-1.75)	941:1379	1.51 (1.22-1.86)
Europe	2211:5862	168:718	0.67 (0.55-0.82)	871:2429	1.07 (0.96-1.20)	1544:2962	1.21 (1.05-1.39)
P for interaction			0.1177		0.2823		0.2075

[1] Pooled ORs were computed using random-effects models, study-specific ORs were adjusted, when available, for sex, age, race/ethmicity, social class, alcohol drinking, fruit and vegetable consumption and study center for multicentric studies [2] Reference category
Ca, cases, Co, controls

5. Discussion

Results from our analysis confirm that there is an association between cigarette smoking and gastric cancer risk. A 20% excess risk of gastric cancer was found among current smoker. This risk significantly increase with the number of cigarettes per day to reach 29% of excess risk for smokers of more than 20 cigarettes and, with duration to reach 32% of excess risk for smokers smoking for more than 30 years compared to never smokers. The effect of duration seems to be somewhat stronger in younger subjects and in men. However, this risk declines with times since stopping and seems to reach the level of never cigarette smokers since 10 years after quitting.

These effects of increasing risk are confirm by different statistical models of analysis including linear model and fractional polynomials, considering the number of cigarettes per day and the duration as a continuous variable.

The categorization of the number of smoking cigarettes needs to be interpreted carefully. In almost all cases, cigarette consumption is assessed by self-report. Because cigarettes are packaged in packs of 20, smokers often represent their cigarette consumption around that number or multiples of number of packs per day and the number of cigarette can be biased.

These results are consistent with previous meta-analyses. The OR estimates for current cigarette smokers were slightly lower than that reported in the previous investigations which found an increasing risk comprised between 1.5 and 1.6 [8, 9, 53].

Among strength of the study, the StoP included original and detail data about cigarette smoking for about 10,000 cases and 25,000 controls, which provided a unique opportunity to investigate and quantify accurately the dose- and duration- risk relationships, and among former smokers, the pattern of risk with time since stopping. Our study included a large number of subjects, increasing our ability to examine relationships between risk factor and gastric cancer risk.

Another advantage of our study was detailed information on important covariates, we adjusted on major risk factor of gastric cancer and conducted stratified analysis by demographic factors including sex, age and geographic area. Having information on the main risk factor of gastric cancer, we could lead subanalyses on studies with helicobacter pylori infection information and particularly fixing the effect of HP infection analyzing only controls with a positive infection to HP (all cases being supposed to be infected to HP).

We were able to take into account heterogeneity across study because of appropriate statistical methods and particularly we could confirm our main results thanks to alternative statistical methods to analyze continuous variables.

The relationship between cigarette smoking and gastric cancer may be discerned from the categorical analysis, but the analysis of the variable in continuous through polynomials brought additional information in particular to understand the possible threshold and a possible change in slopes. Moreover, in a pooled analysis heterogeneity across study can brought differences on the meaning and definition of cutpoints choice and interpretation could be biased.

Moreover, fractional polynomials are an approach transparent, informative, flexible and more realistic than estimation using categories.

However if categorization should be done, in particular for continuous confounders but to avoid excessive bias caused residual confounding, at least four categories are needed [54, 55].

The point of this thesis is to adapt fractional polynomials to be used in a two stage analysis. This method has been developed previously by Rota et al. to use specifically in meta-analyses. We decided to adapt it to pooled case-control studies (and it would be the same for individual patient data meta analyses). During my second year of PhD, I considered also restricted cubic spline as a method for studying non-linear relationship. Analyses on aggregated data gave similar results, but splines are more complex in particular in the choice of the number and the value of nodes.

To our knowledge all studies on individual patient data meta-analysis which considered continuous variable through spline, fractional polynomials and other models from the GAM family, based their analisis on a one-stage approach. This method consists in fitting a regression model, generally a random effect model, adjusting for confounding variables and the effect of the study [56]. However, the availability of confounding variable can varies across study, and it is the cas in the StoP project, the variable adjustment need to be restricted. For this reason the use of the two-stage analysis in the StoP project is inevitable and also preferable in order to include all the study members of the consortium.

However, when the majority of missing data are the results of non-availability of certain variables in some studies, the use of both multiple imputation and the missing-data indicator could be helpful in order to compare results between a one-stage and two-stage analysis [57]. Nevertheless, It would be likely to introduce a bias in comparison with the complete case-method [58] and a two-stage approach would be prefered[57].

V. CONCLUSION AND PERSPECTIVES

During my PhD, I took part of the international consortium of case-control studies on gastric cancer, the “Stomach cancer Pooling (StoP) Project” and specifically in the study of adequate methods to analyze continuous variables in a pooled case-control studies. During this 3 year, I went 6 months at the Mount Sinai hospital in New York to work with Prof. Paolo Boffetta, I had the chance to be involved in two projects on alcohol and cancer which both allowed me to write two scientific publications.

Since a large harmonized data are available in the StoP project (23 studies), I could apply the statistical methods I studied and learn on tobacco smoking and gastric cancer risk. I analyzed in particular continuous risk variables (number of smoked cigarettes per day and the smoking duration) through the classical approach, categorizing the continuous variable, the linear model and fractional polynomials (first and second order), all methods in a two-stage approach. The three methods led to similar conclusions regarding the association between smoking and gastric cancer risk but polynomials gave additional information in particular to understand the possible threshold and changes in slopes.

Further considerations could be lead in the future, in particular regarding if other flexible methods, *i.e.* spline or methods from the GAM family, bring similar results. Moreover, in order to take advantage of the large number of cases and controls, continuous variables it could be interesting to compare our result to a one stage approach overcoming to missing variables using multiple imputation or missing-data indicator methods [59].

REFERENCE

1. Ferlay, J., et al., *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013.* Available from: http://globocan.iarc.fr, accessed on 23/06/2015, 2013.

2. Bertuccio, P., et al., *Recent patterns in gastric cancer: a global overview.* Int J Cancer, 2009. **125**(3): p. 666-73.

3. La Vecchia, C. and S. Franceschi, *Nutrition and gastric cancer.* Can J Gastroenterol, 2000. **14 Suppl D**: p. 51D-54D.

4. Nyrèn, O. and H.-O. Adami, *Stomach cancer*, in *Cancer Epidemiology, Second Edition* H.-O. Adami, D. Hunter, and D. Trichpoulos, Editors. 2008, Oxford Universiy Press: New York. p. 239-274.

5. World Cancer Research Fund and American Institute for Cancer Research, *Food, Nutrition, Physical Activity and the Prevention of Cancer: a Global Perspective. Washington, DC: AICR.* 2007.

6. Bertuccio, P., et al., *Nutrient dietary patterns and gastric cancer risk in Italy.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(11): p. 2882-6.

7. IARC Working Group, *Tobacco smoking. In: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 100E: Personal Habits and indoor combustion.* International Agency for Reseach on Cancer: Lyon, 2012: p. 43-211.

8. Ladeiras-Lopes, R., et al., *Smoking and gastric cancer: systematic review and meta-analysis of cohort studies.* Cancer Causes Control, 2008. **19**(7): p. 689-701.

9. La Torre, G., et al., *Smoking status and gastric cancer risk: an updated meta-analysis of case-control studies published in the past ten years.* Tumori, 2009. **95**(1): p. 13-22.

10. Peleteiro, B., et al., *Worldwide Burden of Gastric Cancer Attributable to Tobacco Smoking in 2012 and Predictions for 2020.* Dig Dis Sci, 2015. **60**(8): p. 2470-6.

11. Gonzalez, C.A., et al., *Smoking and the risk of gastric cancer in the European Prospective Investigation Into Cancer and Nutrition (EPIC).* Int J Cancer, 2003. **107**(4): p. 629-34.

12. Nomura, A.M., et al., *The association of cigarette smoking with gastric cancer: the multiethnic cohort study.* Cancer Causes Control, 2012. **23**(1): p. 51-8.

13. Tramacere, I., C. La Vecchia, and E. Negri, *Tobacco smoking and esophageal and gastric cardia adenocarcinoma: a meta-analysis.* Epidemiology, 2011. **22**(3): p. 344-9.

14. Koizumi, Y., et al., *Cigarette smoking and the risk of gastric cancer: a pooled analysis of two prospective studies in Japan.* Int J Cancer, 2004. **112**(6): p. 1049-55.

15. Freedman, N.D., et al., *A prospective study of tobacco, alcohol, and the risk of esophageal and gastric cancer subtypes.* Am J Epidemiol, 2007. **165**(12): p. 1424-33.

16. Kim, Y., et al., *[Cigarette smoking and gastric cancer risk in a community-based cohort study in Korea].* J Prev Med Public Health, 2007. **40**(6): p. 467-74.

17. Zendehdel, K., et al., *Risk of gastroesophageal cancer among smokers and users of Scandinavian moist snuff.* Int J Cancer, 2008. **122**(5): p. 1095-9.

18. La Vecchia, C., et al., *A pooled analysis of case-control studies of thyroid cancer. III. Oral contraceptives, menopausal replacement therapy and other female hormones.* Cancer Causes Control, 1999. **10**(2): p. 157-66.

19. Hashibe, M., et al., *Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium.* J Natl Cancer Inst, 2007. **99**(10): p. 777-89.

20. DerSimonian, R. and N. Laird, *Meta-analysis in clinical trials.* Control Clin Trials, 1986. **7**(3): p. 177-88.

21. Stukel, T.A., et al., *Two-stage methods for the analysis of pooled data.* Stat Med, 2001. **20**(14): p. 2115-30.

22. Figueiras, A. and C. Cadarso-Suarez, *Application of nonparametric models for calculating odds ratios and their confidence intervals for continuous exposures.* Am J Epidemiol, 2001. **154**(3): p. 264-75.

23. Rota, M., et al., *Random-effects meta-regression models for studying nonlinear dose-response relationship, with an application to alcohol and esophageal squamous cell carcinoma.* Stat Med, 2010. **29**(26): p. 2679-87.

24. Royston, P. and D.G. Altman, *Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling.* App Statist, 1994. **43**(3): p. 429-67.

25. van Houwelingen, H.C., L.R. Arends, and T. Stijnen, *Advanced methods in meta-analysis: multivariate approach and meta-regression.* Stat Med, 2002. **21**(4): p. 589-624.

26. Royston, P. and W. Sauerbrei, *Multivariable Model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* 2008, Chichester: Wiley.

27. Sauerbrei, W. and P. Royston, *Continuous variables: to categorize or to model?In C. Reading (Ed.), Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.*

28. Mao, Y., et al., *Active and passive smoking and the risk of stomach cancer, by subsite, in Canada.* Eur J Cancer Prev, 2002. **11**(1): p. 27-38.

29. Zhang, Z.F., et al., *Adenocarcinomas of the esophagus and gastric cardia: medical conditions, tobacco, alcohol, and socioeconomic factors.* Cancer Epidemiol Biomarkers Prev, 1996. **5**(10): p. 761-8.

30. Trichopoulos, D., et al., *Diet and cancer of the stomach: a case-control study in Greece.* Int J Cancer, 1985. **36**(3): p. 291-7.

31. Augustin, L.S., et al., *Glycemic index, glycemic load and risk of gastric cancer.* Ann Oncol, 2004. **15**(4): p. 581-4.

32. Lucenteforte, E., et al., *Food groups and alcoholic beverages and the risk of stomach cancer: a case-control study in Italy.* Nutr Cancer, 2008. **60**(5): p. 577-84.

33. De Feo, E., et al., *A case-control study on the effect of Apolipoprotein E genotypes on gastric cancer risk and progression.* BMC Cancer, 2012. **12**: p. 494.

34. Buiatti, E., et al., *A case-control study of gastric cancer and diet in Italy.* Int J Cancer, 1989. **44**(4): p. 611-6.

35. Lunet, N., et al., *Antioxidant vitamins and risk of gastric cancer: a case-control study in Portugal.* Nutr Cancer, 2006. **55**(1): p. 71-7.

36. Zaridze, D., et al., *Aspirin protects against gastric cancer: results of a case-control study from Moscow, Russia.* Int J Cancer, 1999. **82**(4): p. 473-6.

37. Santibanez, M., et al., *Occupational exposures and risk of stomach cancer by histological type.* Occup Environ Med, 2012. **69**(4): p. 268-75.

38. Ekstrom, A.M., et al., *Occupational exposures and risk of gastric cancer in a population-based case-control study.* Cancer Res, 1999. **59**(23): p. 5932-7.

39. Deandrea, S., et al., *Is temperature an effect modifier of the association between green tea intake and gastric cancer risk?* Eur J Cancer Prev, 2010. **19**(1): p. 18-22.

40. Mu, L.N., et al., *Green tea drinking and multigenetic index on the risk of stomach cancer in a Chinese population.* Int J Cancer, 2005. **116**(6): p. 972-83.

41. Yu, G.P., et al., *Green-tea consumption and risk of stomach cancer: a population-based case-control study in Shanghai, China.* Cancer Causes Control, 1995. **6**(6): p. 532-8.

42. Setiawan, V.W., et al., *GSTT1 and GSTM1 null genotypes and the risk of gastric cancer: a case-control study in a Chinese population.* Cancer Epidemiol Biomarkers Prev, 2000. **9**(1): p. 73-80.

43. Pourfarzi, F., et al., *The role of diet and other environmental factors in the causation of gastric cancer in Iran--a population based study.* Int J Cancer, 2009. **125**(8): p. 1953-60.

44. Pakseresht, M., et al., *Dietary habits and gastric cancer risk in north-west Iran.* Cancer Causes Control, 2011. **22**(5): p. 725-36.

45. Derakhshan, M.H., et al., *Combination of gastric atrophy, reflux symptoms and histological subtype indicates two distinct aetiologies of gastric cardia cancer.* Gut, 2008. **57**(3): p. 298-305.

46. Inoue, M., et al., *Epidemiological features of first-visit outpatients in Japan: comparison with general population and variation by sex, age, and season.* J Clin Epidemiol, 1997. **50**(1): p. 69-77.

47. Lopez-Carrillo, L., et al., *Nutrient intake and gastric cancer in Mexico.* Int J Cancer, 1999. **83**(5): p. 601-5.

48. Lopez-Carrillo, L., M. Hernandez Avila, and R. Dubrow, *Chili pepper consumption and gastric cancer in Mexico: a case-control study.* Am J Epidemiol, 1994. **139**(3): p. 263-71.

49. Lopez-Carrillo, L., et al., *Capsaicin consumption, Helicobacter pylori positivity and gastric cancer in Mexico.* Int J Cancer, 2003. **106**(2): p. 277-82.

50. Hamada, G.S., et al., *Risk factors for stomach cancer in Brazil (II): a case-control study among Japanese Brazilians in Sao Paulo.* Jpn J Clin Oncol, 2002. **32**(8): p. 284-90.

51. Machida-Montani, A., et al., *Association of Helicobacter pylori infection and environmental factors in non-cardia gastric cancer in Japan.* Gastric Cancer, 2004. **7**(1): p. 46-53.

52. Smith-Warner, S.A., et al., *Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer.* Am J Epidemiol, 2006. **163**(11): p. 1053-64.

53. Tredaniel, J., et al., *Tobacco smoking and gastric cancer: review and meta-analysis.* Int J Cancer, 1997. **72**(4): p. 565-73.

54. Becher, H., *The concept of residual confounding in regression models and some applications.* Stat Med, 1992. **11**(13): p. 1747-58.

55. Brenner, H. and M. Blettner, *Controlling for continuous confounders in epidemiologic research.* Epidemiology, 1997. **8**(4): p. 429-34.

56. Riley, R.D., *Commentary: like it and lump it? Meta-analysis using individual participant data.* Int J Epidemiol, 2010. **39**(5): p. 1359-61.

57. Raimondi, S., et al., *Melanocortin-1 receptor, skin cancer and phenotypic characteristics (M-SKIP) project: study design and methods for pooling results of genetic epidemiological studies.* BMC Med Res Methodol, 2012. **12**: p. 116.

58. Huberman, M. and B. Langholz, *Application of the missing-indicator method in matched case-control studies with incomplete data.* Am J Epidemiol, 1999. **150**(12): p. 1340-5.

59. Jolani, S., et al., *Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE.* Stat Med, 2015. **34**(11): p. 1841-63.

SUPPLEMENTS

1. SAS Macro for fractional polynomials

```
/********************************************************************
                              MACRO MFP

Macro that build the dataset in order to fit first order fractional
polynomials

Input parameters->
powers: vector containing the list of powers we want to try in our
fractional polynomials vector
data:   dataset containing our dataset
var:    dependant continuous variable

********************************************************************/
%macro MFP1(powers,data,var,num_studies);

*** count the number of different powers to test ;
proc iml;
 a=ncol({&powers}); /* counter of the number of power in the vector
*/
 create tt from a[colname="n_powers"];  /* create a dataset which
contained a */
 append from a;
data s;
 set tt;
 call symput('cont',n_powers);
run;
quit;
*** create a dataset with the value of p1 in funzione of its rank in
the vector;
%do j=1 %to &cont;       /* counter of the power p1 */
  %let a=%qscan(&powers,&j,%str( ));
     data new;
       set &data;
         %if &a ne 0 %then %do;   /* p1 different from 0 */
        p1=&var**(&a); %end;
         %else %do;                      /* p1=0*/
        p1=log(&var); %end;
       run;
       %logistic1(new,&a,&j,&risposta,&num_studies);
%end;
```

```
/********************************************************************
                            MACRO LOGISTIC

OGGETTO: fit the logistic regression on the variable for each power
Input parameters->
dataset: dataset containing the transformed variable with the power
p1:      the transformed explicative variable
index:   index which for the 8 polynomials
y:       represent the explained variable
num_studies: the number refered to the last study considered
********************************************************************/


options symbolgen mprint;
%macro logistic1(dataset,p1,index,y,num_studies);

%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));

      %let stud=%scan(&studies,&nstudy," ");

*** create a dataset for each study;
      data st&stud;
      set &dataset;
      if va2=&stud. and va2 in (&studies);
      run;

*** compute the logistic model for each study and for each power
taking out the beta estimate and the coresponding variance;
                  title "Study &stud.";
                  proc logistic data=st&stud;
                  class &&study&stud.  / ref=FIRST param=ref;
                  model &y=p1  &&study&stud. / link=logit covb;
                  *where va2=&stud.;
                  ods output ParameterEstimates=StimaBeta&stud
CovB=StimaCov&stud /*FitStatistics=Fit_for_label&i*/;
                  run;
                  quit;

*** create a dataset with betas for each power;
            data StimaBeta&stud;
            set StimaBeta&stud;
            powers_1=&p1;
            va2=&stud;
            if Variable eq "p1";
            drop ClassVal0 DF WaldChiSq ProbChiSq;
            run;
            quit;

*** create a dataset with the variances for each power;
            data StimaCov&stud;
            set StimaCov&stud;
            if Parameter eq "p1" ;
            va2=&stud;
            rename p1=var_p1;
            keep p1 va2;
            run;
      %let nstudy=%eval(&nstudy+1);
%end;
```

```
*** Create a dataset merging beta estimates of each study together
with each power;
data beta;
set
%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));
      %let stud=%scan(&studies,&nstudy," ");
      Stimabeta&stud
      %let nstudy=%eval(&nstudy+1);
      %end;
;
b=1;
run;

*** Create a dataset merging variance estimates of each study
together with each power;
data covb;
set
%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));
      %let stud=%scan(&studies,&nstudy," ");
      Stimacov&stud
      %let nstudy=%eval(&nstudy+1);
      %end;
run;


*** Create the dataset "longformat" with the variances in the format
needed for the proc mixed;
proc transpose data=covb out=w(drop=_name_);
 var var_p1;
run;

proc iml;
use w;
read all into d;
d={0}||d;
nc=ncol(d);
num=1:nc;
colname=cat('CovP',num);
create longformat from d [colname=colname];
append from d;

quit;

*** Estimate the pooled beta for each power.
M1-M8 datasets contain the pooled beta for each power.
F1-F8 datasets ccontain -2logV AIC ... for each power;
title "Pooled";
proc mixed data=beta covtest cl;
 class va2;
 model estimate=b /noint solution cl covb;
 random b /subject=va2 g solution type=un;
 repeated /subject=va2 group=va2 type=un;
 parms/parmsdata=longformat hold=2 to &num_studies.+1;
 ods output SolutionF=m&index FitStatistics=f&index;
run;quit;
```

```
data m&index;
set m&index;
 powers_1=&p1;
run;


* F1-F8 datasets ccontain -2logV for each power;
data f&index.verosim;
set f&index;
if Descr eq "-2 res log verosim";
*if Descr eq "-2 Res Log Likelihood";
powers_1=&p1;
run;

* F1-F8 datasets ccontain AIC for each power;
data f&index.aic;
set f&index;
if Descr eq "AIC (minore è meglio)";
*if Descr eq "AIC (smaller is better)";
powers_1=&p1;
run;

* create an only one dataset merging F1-F8 with -2logV;
data fitverosim;
set
%do l=1 %to &cont;
    f&l.verosim
%end;
run;

* create an only one dataset merging F1-F8 with AIC;
data fitaic;
set
%do l=1 %to &cont;
    f&l.aic
%end;
run;

* create an only one dataset with beta estimates for each power;
data coeff1;
set
%do l=1 %to &cont;
    m&l
%end;
run;
%mend;
```

```
/**********************************************************************
                              MACRO MFP2

Macro that build the in order to fit second order fractional
polynomials

Input parameters->
powers: vector containing the list of powers we want to try in our
fractional polynomials vector
data:   dataset containing our data
var:    dependant continuous variable
num_studies: the number refered to the last study considered
**********************************************************************/

%macro MFP2(powers,data,var,num_studies);


proc iml;
a=ncol({&powers}); /* counter of the number of power in the vector */
create tt from a[colname="n_powers"];  /* create a dataset which
contained a */
append from a;
data s;
set tt;
call symput('cont',n_powers);
run;
quit;

%do j=1 %to &cont;    /* counter of the power p1 */
%do k=1 %to &cont;    /* counter of the power p2 */

      %if &j=&k %then %do; /* p1=p2 */
            %let a=%qscan(&powers,&j,%str( ));
            data new;
            set &data;
            %if &a ne 0 %then %do;   /* p1=p2 different from 0 */
                  p1=&var**(&a);
                  p2=&var**(&a)*log(&var);
                  %end;
            %else %do;                    /* p1=p2=0*/
                  p1=log(&var);
                  p2=(log(&var))**2;
                  %end;
            run;

            %logistic2(new,&a,&a,&k&j,&risposta,&num_studies);

            %end;


      %else %do; /* p1 not equal to p2 */
            %let a=%qscan(&powers,&j,%str( ));
            %let b=%qscan(&powers,&k,%str( ));
            data new;
            set &data;
            %if &a ne 0 AND &b ne 0 %then %do; /* p1 and p2 not equal
to 0 */
                  p1=&var**(&a);
                  p2=&var**(&b);
                  %end;
            %else %if &a eq 0 AND &b ne 0 %then %do; /* p1=0 but p2
not equal to 0 */
                  p1=log(&var);
                  p2=&var**(&b);
                  %end;
            %else %do;
                  p1=&var**(&a);
                  p2=log(&var);
                  %end;
            run;
            %logistic2(new,&a,&b,&j&k,&risposta,&num_studies);

            %end;
%end;
%end;

%mend;


/********************************************************************
                           MACRO LOGISTIC2

OGGETTO: fit the logistic regression on the variable for each power

Input parameters->
dataset:    dataset containing the transformed variable with the
power
p1:         the transformed explicative variable
p2:         the transformed explicative variable
index:      index which for the 36 polynomials
y:                represent the explained variable
num_studies: the number refered to the last study considered
********************************************************************/
options symbolgen mprint;
%macro logistic2(dataset,p1,p2,index,y,num_studies);

%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));

%let stud=%scan(&studies,&nstudy," ");

*** create a dataset for each study;
data st&stud;
set &dataset;
if va2=&stud. and va2 in (&studies);
run;

*** compute the logistic model for each study and for each couple of
powers taking out the beta estimate and the coresponding variance;
            title "Study &stud.";
            proc logistic data=st&stud;
            class &&study&stud.  / ref=FIRST param=ref;
            model &y=p1 p2  &&study&stud. / link=logit covb;
            ods output ParameterEstimates=StimaBeta&stud
CovB=StimaCov&stud ;
            run;
            quit;
```

```
*** create a dataset with betas for each power;
data StimaBeta&stud;
set  StimaBeta&stud;
powers_1=&p1;
powers_2=&p2;
va2=&stud;
if Variable = "p1" or variable="p2" then output;
drop ClassVal0 DF WaldChiSq ProbChiSq;
run;
quit;

*** create a dataset with the variances for each power;
data StimaCov&stud;
set  StimaCov&stud;
if Parameter = "p1" then do; p1bis=p1;end;
var_p1st&stud=lag1(p1bis);
if Parameter="p2" then do;cov_p1p2st&stud=p1; var_p2st&stud=p2;end;
if Parameter="p2" then output;
keep var_p1st&stud var_p2st&stud cov_p1p2st&stud;
run;
%let nstudy=%eval(&nstudy+1);
%end;

*** Create a dataset merging beta estimates of each study together
with each power;
data beta;
set
%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));
      %let stud=%scan(&studies,&nstudy," ");
      Stimabeta&stud
      %let nstudy=%eval(&nstudy+1);
      %end;
;

if variable="p1" then b1=1;else b1=0;
if variable="p2" then b2=1;else b2=0;

run;

*** Create a dataset merging variance estimates of each study
together with each power;
data covb;
merge
%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));
      %let stud=%scan(&studies,&nstudy," ");
      Stimacov&stud
      %let nstudy=%eval(&nstudy+1);
      %end;
;
run;


*** Create the dataset "longformat" with the variances in the format
needed for the proc mixed;
proc iml;
use covb;
```

```
read all into d;
d={0}||{0}||{0}||d;
nc=ncol(d);
num=1:nc;
colname=cat('CovP',num);
create longformat from d [colname=colname];
append from d;

quit;


data _null_;
call symput('num_col',&num_studies*3+3);
run;

*** Estimate the pooled beta for each power.
M1-M88 datasets contain the pooled beta for each couple of power.
F1-F888 datasets ccontain -2logV AIC ... for each couple of power;
title "Pooled";
ods graphics on;
proc mixed data=beta covtest cl;
 class va2;
 model estimate=b1 b2 /noint solution cl covb;
 random b1 b2 /subject=va2 g solution type=un;
 repeated /subject=va2 group=va2 type=un;
 parms/parmsdata=longformat hold=4 to &num_col;
 ods output SolutionF=m&index FitStatistics=f&index;
run;
quit;
ods graphics off;

data m&index;
set m&index;
powers_1=&p1;
powers_2=&p2;
run;

data f&index.aic;
set f&index;
if Descr eq "AIC (minore è meglio)";
powers_1=&p1;
powers_2=&p2;
run;

data f&index.verosim;
set f&index;
if Descr eq "-2 res log verosim";
powers_1=&p1;
powers_2=&p2;
run;

data fitaic;
set
%do l=1 %to &cont;
%do m=1 %to &cont;
     f&l&m.aic
%end;
%end;
;
```

```
run;
data fitverosim;
set
%do l=1 %to &cont;
%do m=1 %to &cont;
     f&l&m.verosim
%end;
%end;
;
run;
proc sort data=fitaic;by Value; run;
proc sort data=fitverosim;by Value; run;

* create an only one dataset with beta estimates for each power;
data coeff2;
set
%do l=1 %to &cont;
%do m=1 %to &cont;
      m&l&m
%end;
%end;
;
run;

%mend;
```

2. R programs for fractional polynomials

Example of program to build the graphic

```
#second order polynomials
eta1<- -0.00002
beta2<- 0.000022

numcig<-seq(0.01,100,1)
a<-numcig^-2
b<-numcig^2

matx<-cbind(a,b)

predicted<-exp(beta1*a+beta2*b)
cov<-matrix(nrow=2,ncol=2)

cov[1,1]<-  2.07E-11
cov[1,2]<-  -212E-13
cov[2,1]<-cov[1,2]
cov[2,2]<-  8.48E-10

lb_predicted<-exp((beta1*a+beta2*b)-1.96*sqrt(diag(matx%*%cov%*%t(matx))))
ub_predicted<-exp((beta1*a+beta2*b)+1.96*sqrt(diag(matx%*%cov%*%t(matx))))
predicted[1]<-1
lb_predicted[1]<-1
```

```
ub_predicted[1]<-1
predicted

#linear trend
beta<- 0.007640
var<- 0.001136
predicted_lin<-exp(beta*numcig)

lb_predicted_lin<-exp((beta*numcig)-1.96*sqrt(numcig*var*t(numcig)))
ub_predicted_lin<-exp((beta*numcig)+1.96*sqrt(numcig*var*t(numcig)))
summary(numcig)
predicted

# graph FP e linear together
plot(numcig,predicted,type="l",xlab="Number of smoking cigarettes per
day",ylab="OR",xlim=c(0,100),ylim=c(0.5,4),log="y",yaxt="n")
axis(2,at=c(0.5,1,2,4),las=1)

polygon(c(numcig, rev(numcig)), c(ub_predicted, rev(lb_predicted)),
        col = "gray87", border = NA)
lines(numcig,lb_predicted, type='l',lty=3)
lines(numcig,ub_predicted, type='l',lty=3)
lines(numcig,predicted, type='l',lwd=2)
lines(numcig,predicted_lin,col="black",lty=5)
legend(28,4.4, c("Linear model", "Fractional polynomials"),
col=c("black","black"),lty=c(5,1),lwd=c(1,2),bty="n")
```

3. SAS programs for analysis of the 2 stage method: first step

```
/*********************************************************************
                            MACRO TWO STEP
dataset -> dataset to analyse
studies -> macrovariable with the studies considered
case_control -> outcome variable (va1 case/control)
exposure -> exposure variable
exp_ref_cat -> reference category
strata -> strata variable (if I don't want to conduct a stratified
analysis, put "")
format_exp -> exposure variable format
num_studies -> the number of the last study considered
*********************************************************************/


%macro twostep
(dataset,studies,case_control,exposure,exp_ref_cat,strata,format_exp,
num_studies);

ods trace on;
```

```
%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));

 %let stud=%scan(&studies,&nstudy," ");

 %if &strata eq "" %then %do;
 *** for each study a logistic regression is performed with its
specific adjustment;
 proc logistic data=&dataset ;
  title "Study &stud.";
  format &exposure ;
  class &exposure (ref="&exp_ref_cat") &&study&stud / param=ref;
  model &case_control= &exposure &&study&stud ;
  where va2=&stud;
  ods output ParameterEstimates=beta_study&stud; /* output of beta
estimates */
 run;
 quit;
 %end;
 %else %do;

 *** in strata of the strata variable;
  proc sort data=&dataset;by &strata;run;
  proc logistic data=&dataset ;
  title "Study &stud.";
  format &exposure ;
  class &exposure (ref="&exp_ref_cat") &&study&stud / param=ref;
  model &case_control= &exposure &&study&stud ;
  by &strata;
  where va2=&stud;
  ods output ParameterEstimates=beta_study&stud; /* output of beta
estimates */
 run;
 quit;
 %end;

 *** We created a dataset with the only variables we need for the
pooled analysis;
 data beta&stud;
  set beta_study&stud;

  study=&stud;

  if variable eq "&exposure";

  &exposure=input(ClassVal0,best12.);
  format &exposure &format_exp..;

   label &exposure="&exposure";

   drop df WaldChiSq ProbChiSq variable ClassVal0;
 run;

 %if &strata eq "" %then %do;
 proc freq data=&dataset noprint;
  table &exposure * &case_control /out=freqs&stud;
  where va2=&stud;
 run; /* nota. The proc  freq allow to have the number of cases and
controls in each strata in order to be used in the forest plot */
```

```
 %end;
 %else %do;
 proc freq data=&dataset noprint;
  table &exposure * &case_control /out=freqs&stud;
  by &strata;
  where va2=&stud;
 run;
 %end;

 data freqs&stud;
  set freqs&stud;

  study=&stud;

  format &exposure &format_exp..;
 run;

%if &strata eq "" %then %do;
proc sort data=freqs&stud;by study &exposure;run;
 proc transpose data=freqs&stud out=t_freq&stud (drop=_NAME_ _LABEL_
rename=(COL1=CASES COL2=CONTROLS));
  var count;
  by study &exposure;
  where &exposure ne .a and &exposure ne .c ;
 run;
%end;
%else %do;
 proc sort data=freqs&stud;by study &exposure &strata;run;
 proc transpose data=freqs&stud out=t_freq&stud (drop=_NAME_ _LABEL_
rename=(COL1=CASES COL2=CONTROLS));
  var count;
  by study &exposure &strata;
  where &exposure ne .a and &exposure ne .c ;
 run;
%end;


 %let nstudy=%eval(&nstudy+1);

%end;

ods trace off;

data beta_all_studies;
 set

 %let nstudy=1;
 %do %while (%length(%scan(&studies,&nstudy," ")));

 %let stud=%scan(&studies,&nstudy," ");
   beta&stud

   %let nstudy=%eval(&nstudy+1);
 %end;
 ;

run;

data freqs_all_studies;
```

```
 set

 %let nstudy=1;
 %do %while (%length(%scan(&studies,&nstudy," ")));
 %let stud=%scan(&studies,&nstudy," ");


   t_freq&stud

 %let nstudy=%eval(&nstudy+1);
 %end;

 ;

  where study ne . and (CASES ne . AND CONTROLS ne .);
run;

proc datasets lib=work nolist;
 delete Beta_study1-Beta_study&num_studies Beta1-Beta&num_studies
freqs1-freqs&num_studies t_freq1-t_freq&num_studies;
quit;

%if &strata eq "" %then %do;
proc sort data=beta_all_studies;by study &exposure;run;
proc sort data=freqs_all_studies;by study &exposure;run;

data &exposure;
 merge beta_all_studies (in=a) freqs_all_studies (in=b) ;
 if a and b;
 by study &exposure;
 format study $studyb.;
run;



PROC EXPORT DATA= &exposure
            OUTFILE= "&dir.&exposure..csv"
            DBMS=CSV REPLACE;
RUN;

%end;
%else %do;
proc sort data=beta_all_studies;by study &exposure &strata;run;
proc sort data=freqs_all_studies;by study &exposure &strata;run;

data &exposure._&strata;
 merge beta_all_studies (in=a) freqs_all_studies (in=b);
 if a and b;
 by study &exposure &strata;
 format study $study.;
run;

PROC EXPORT DATA= &exposure._&strata
            OUTFILE= "&dir.&Exposure._&strata..csv"
            DBMS=CSV REPLACE;
RUN;

%end;%mend;
```

```
/*********************************************************************
                  MACRO TWO STEP to alyse the TREND
dataset -> dataset to analyse
studies -> macrovariable with the studies considered
case_control -> outcome variable (va1 case/control)
exposure -> exposure variable
exp_ref_cat -> reference category
strata -> strata variable (if I don't want to conduct a stratified
analysis, put "")
format_exp -> exposure variable format
num_studies -> the number of the last study considered
*********************************************************************/



%macro twosteptr
(dataset,studies,case_control,exposure,exp_ref_cat,strata,format_exp,
num_studies);

ods trace on;

%let nstudy=1;
%do %while (%length(%scan(&studies,&nstudy," ")));

 %let stud=%scan(&studies,&nstudy," ");

 %if &strata eq "" %then %do;
 proc logistic data=&dataset ;
  title "Study &stud.";
  format &exposure ;
  class /*&exposure (ref="&exp_ref_cat")*/ &&study&stud / param=ref;
  model &case_control= &exposure &&study&stud ;
  where va2=&stud;
  ods output ParameterEstimates=beta_study&stud; /* output of beta
estimates */
 run;
 quit;
 %end;
 %else %do;
  proc sort data=&dataset;by &strata;run;
  proc logistic data=&dataset ;
  title "Study &stud.";
  format &exposure ;
  class /*&exposure (ref="&exp_ref_cat")*/ &&study&stud / param=ref;
  model &case_control= &exposure &&study&stud ;
  by &strata;
  where va2=&stud;
  ods output ParameterEstimates=beta_study&stud; /* output of beta
estimates */
 run;
 quit;
 %end;

 data beta&stud;
  set beta_study&stud;

  study=&stud;
  if variable eq "&exposure";

  &exposure=input(ClassVal0,best12.);
  format &exposure &format_exp..;

   label &exposure="&exposure";

   drop df WaldChiSq ProbChiSq variable ClassVal0;
 run;

 %if &strata eq "" %then %do;
 proc freq data=&dataset noprint;
  table /*&exposure **/ &case_control /out=freqs&stud;
  where va2=&stud and &exposure ne .;
 run; /* nota. La proc freq mi serve per tabulare numero di casi e
controlli entro ogni strato per il forest plot */
 %end;
 %else %do;
 proc freq data=&dataset noprint;
  table /* &exposure **/ &case_control /out=freqs&stud;
  by &strata;
  where va2=&stud and &exposure ne .;
 run;
 %end;

 data freqs&stud;
  set freqs&stud;

  study=&stud;

  format &exposure &format_exp..;
 run;

%if &strata eq "" %then %do;
proc sort data=freqs&stud;by study &exposure;run;
 proc transpose data=freqs&stud out=t_freq&stud (drop=_NAME_ _LABEL_
rename=(COL1=CASES COL2=CONTROLS));
  var count;
  by study &exposure;
  where &exposure ne .a and &exposure ne .c ;
 run;
%end;
%else %do;
 proc sort data=freqs&stud;by study &exposure &strata;run;
 proc transpose data=freqs&stud out=t_freq&stud (drop=_NAME_ _LABEL_
rename=(COL1=CASES COL2=CONTROLS));
  var count;
  by study &exposure &strata;
  where &exposure ne .a and &exposure ne .c ;
 run;
%end;


 %let nstudy=%eval(&nstudy+1);

%end;

ods trace off;
```

```
data beta_all_studies;
 set

 %let nstudy=1;
 %do %while (%length(%scan(&studies,&nstudy," ")));

 %let stud=%scan(&studies,&nstudy," ");

   beta&stud

 %let nstudy=%eval(&nstudy+1);
 %end;

 ;

run;

data freqs_all_studies;
 set

 %let nstudy=1;
 %do %while (%length(%scan(&studies,&nstudy," ")));
 %let stud=%scan(&studies,&nstudy," ");


   t_freq&stud

%let nstudy=%eval(&nstudy+1);
 %end;

 ;

  where study ne . and (CASES ne . AND CONTROLS ne .);
run;

proc datasets lib=work nolist;
 delete Beta_study1-Beta_study&num_studies Beta1-Beta&num_studies
freqs1-freqs&num_studies t_freq1-t_freq&num_studies;
quit;


%if &strata eq "" %then %do;
proc sort data=beta_all_studies;by study &exposure;run;
proc sort data=freqs_all_studies;by study &exposure;run;

data &exposure;
 merge beta_all_studies (in=a) freqs_all_studies (in=b);
 if a and b;
 by study &exposure;
 format study $study.;
run;

PROC EXPORT DATA= &exposure
            OUTFILE= "&dir.&exposure.tr.csv"
            DBMS=CSV REPLACE;
RUN;

%end;
%else %do;
```

```
proc sort data=beta_all_studies;by study &exposure &strata;run;
proc sort data=freqs_all_studies;by study &exposure &strata;run;

data &exposure._&strata;
 merge beta_all_studies (in=a) freqs_all_studies (in=b);
 if a and b;
 by study &exposure &strata;
 format study $study.;
run;

PROC EXPORT DATA= &exposure._&strata
            OUTFILE= "&dir.&Exposure._&strata.tr.csv"
            DBMS=CSV REPLACE;
RUN;

%end;

%mend;
```

4. R programs for analysis of the 2 stage method: second step

```
cigday<-read.csv(file="F:\\DOTTORATO\\RELAZIONE TERZO
ANNO\\SAS\\SMOKING\\DATASETS FOR R\\cigarette_dayarea.csv")
attach(cigday)

library(meta)
# cigday
# 1="Never smoker"
# 2="Former cigarette smoker"
# 3="0 to 10"
# 4="11 to 20"
# 5="> 20"
# 6="Other than cigarette smoker";

# ALL STRATA INTO ONE GRAPH
me<-metagen(TE=Estimate[cigarette_day!="Never smoker"],
seTE=StdErr[cigarette_day!="Never smoker"], studlab=study[cigarette_day!="Never
smoker"], sm="OR",n.e=CASES[cigarette_day!="Never smoker"],
n.c=CONTROLS[cigarette_day!="Never smoker"],
byvar=cigarette_day[cigarette_day!="Never smoker"],level=0.95,level.comb=0.95,
comb.random=TRUE,comb.fixed=FALSE,method.tau="DL")

setwd("F:\\DOTTORATO\\RELAZIONE TERZO ANNO\\RESULTS\\Smoking
forest plot pdf\\")
pdf(file="cigarette_day.pdf",paper = "a4", width = 11, height = 11,pagecentre=T,
pointsize=7)
```

```
forest.meta(me,pooled.totals=FALSE,pooled.events=TRUE,smlab="",xlab="Odds
Ratio",ref=1,overall=TRUE,print.I2=T,leftlabs=c("Study","Cancer
cases","Controls","OR","95% CI"),xlim=c(0.25,4),col.square="black",col.by="black",
addspace=TRUE,print.tau2=FALSE,rightcols=FALSE,leftcols=c("studlab","n.e","n.c"
,"effect","ci"),print.byvar=FALSE,text.random="Pooled estimate",
ff.random.labels=0.5,ff.random=0.5,ff.hetstat=0.5,fontsize=8,squaresize=1)
dev.off()


# Subset: Former cigarette smoker

me<-metagen(TE=Estimate[cigarette_day=="Former cigarette smoker"],
seTE=StdErr[cigarette_day=="Former cigarette smoker"],
studlab=study[cigarette_day=="Former cigarette smoker"], sm="OR",
n.e=CASES[cigarette_day=="Former cigarette smoker"],
n.c=CONTROLS[cigarette_day=="Former cigarette smoker"],
byvar=area[cigarette_day=="Former cigarette smoker"], level=0.95,level.comb=0.95,
comb.random=TRUE,comb.fixed=FALSE,method.tau="DL", title="Former cigarette
smoker")

setwd("F:\\DOTTORATO\\RELAZIONE TERZO ANNO\\RESULTS\\Smoking
forest plot pdf\\")
pdf(file="cigarette_day 1- Former cigarette smokertry.pdf", paper = "a4r",  width = 10,
height = 10, pagecentre=T,  pointsize=10)
forest.meta(me, pooled.totals=TRUE, pooled.events=TRUE, smlab=" ", xlab="Odds
Ratio", ref=1, overall=TRUE, print.I2=T, leftlabs=c("Study", "Cancer cases",
"Controls", "OR", "95% CI"), xlim=c(0.25, 4), print.byvar=F, col.square="black",
col.by="black", addspace=TRUE, print.tau2=FALSE, rightcols=FALSE,
leftcols=c("studlab", "n.e", "n.c", "effect", "ci"), text.random="Pooled estimate",
ff.random.labels=2, sortvar=me$studlab, ff.random=1, ff.hetstat=1, fontsize=11,
squaresize=1)
dev.off()
tiff(filename = "F:\\DOTTORATO\\RELAZIONE TERZO
ANNO\\RESULTS\\Smoking forest plot tiff\\cigarette_day 1- Former cigarette
smokerAREA.tiff",   width = 230,  height = 230,  units = "mm",  res=300)
forest.meta(me, pooled.totals=TRUE, pooled.events=TRUE, smlab=" ", xlab="Odds
Ratio", ref=1, overall=TRUE, print.I2=T,  leftlabs=c("Study", "Cancer cases",
"Controls", "OR", "95% CI"), xlim=c(0.25, 4), print.byvar=F,  col.square="black",
col.by="black", addspace=TRUE, print.tau2=FALSE, rightcols=FALSE,
leftcols=c("studlab", "n.e", "n.c", "effect", "ci"), text.random="Pooled estimate",
ff.random.labels=2, sortvar=me$studlab,  ff.random=1, ff.hetstat=1, fontsize=11,
squaresize=1)
dev.off()

# Subset: 0 to 10

me<-metagen(TE=Estimate[cigarette_day=="0 to 10"],
seTE=StdErr[cigarette_day=="0 to 10"], studlab=study[cigarette_day=="0 to 10"],
sm="OR", n.e=CASES[cigarette_day=="0 to 10"],
n.c=CONTROLS[cigarette_day=="0 to 10"], byvar=area[cigarette_day=="0 to 10"],
level=0.95, level.comb=0.95, comb.random=TRUE, comb.fixed=FALSE,
method.tau="DL", title="0 to 10")

setwd("F:\\DOTTORATO\\RELAZIONE TERZO ANNO\\RESULTS\\Smoking
forest plot pdf\\")
pdf(file="cigarette_day 2- 0 to 10.pdf", paper = "a4r",  width = 10,  height = 10,
pagecentre=T,  pointsize=10)
forest.meta(me, pooled.totals=TRUE, pooled.events=TRUE, smlab=" ", xlab="Odds
Ratio", ref=1, overall=TRUE, print.I2=T, leftlabs=c("Study", "Cancer cases",
"Controls", "OR", "95% CI"), xlim=c(0.25, 4), print.byvar=F, col.square="black",
col.by="black", addspace=TRUE, print.tau2=FALSE, rightcols=FALSE,
leftcols=c("studlab", "n.e", "n.c", "effect", "ci"), text.random="Pooled estimate",
ff.random.labels=2, ff.random=1, ff.hetstat=1, fontsize=11, squaresize=1)
dev.off()

tiff(filename = "F:\\DOTTORATO\\RELAZIONE TERZO
ANNO\\RESULTS\\Smoking forest plot tiff\\cigarette_day 2- 0 to 10AREA.tiff",
width = 230,  height = 230,  units = "mm",  res=300)
forest.meta(me, pooled.totals=TRUE, pooled.events=TRUE, smlab=" ", xlab="Odds
Ratio", ref=1, overall=TRUE, print.I2=T, leftlabs=c("Study", "Cancer cases",
"Controls", "OR", "95% CI"), xlim=c(0.25, 4), print.byvar=F, col.square="black",
col.by="black", addspace=TRUE, print.tau2=FALSE, rightcols=FALSE,
leftcols=c("studlab", "n.e", "n.c", "effect", "ci"), text.random="Pooled estimate",
ff.random.labels=2, ff.random=1, ff.hetstat=1, fontsize=11, squaresize=1)
dev.off()

# Subset: 11 to 20

me<-metagen(TE=Estimate[cigarette_day=="11 to 20"],
seTE=StdErr[cigarette_day=="11 to 20"], studlab=study[cigarette_day=="11 to 20"],
sm="OR", n.e=CASES[cigarette_day=="11 to 20"],
n.c=CONTROLS[cigarette_day=="11 to 20"], byvar=area[cigarette_day=="11 to 20"],
level=0.95, level.comb=0.95, comb.random=TRUE, comb.fixed=FALSE,
method.tau="DL", title="11 to 20")

setwd("F:\\DOTTORATO\\RELAZIONE TERZO ANNO\\RESULTS\\Smoking
forest plot pdf\\")
pdf(file="cigarette_day 3- 11 to 20.pdf", paper = "a4r",  width = 10,  height = 10,
pagecentre=T,  pointsize=10)
forest.meta(me, pooled.totals=TRUE, pooled.events=TRUE, smlab=" ", xlab="Odds
Ratio", ref=1, overall=TRUE, print.I2=T, leftlabs=c("Study", "Cancer cases",
"Controls", "OR", "95% CI"), xlim=c(0.25, 4), print.byvar=F, col.square="black",
col.by="black", addspace=TRUE, print.tau2=FALSE, rightcols=FALSE,
leftcols=c("studlab", "n.e", "n.c", "effect", "ci"), text.random="Pooled estimate",
ff.random.labels=2,  ff.random=1, ff.hetstat=1, fontsize=11, squaresize=1)
dev.off()
```

```
tiff(filename = "F:\\DOTTORATO\\RELAZIONE TERZO
ANNO\\RESULTS\\Smoking forest plot tiff\\cigarette_day 3- 11 to 20AREA.tiff",
width = 230,  height = 230,  units = "mm",  res=300)
forest.meta(me, pooled.totals=TRUE, pooled.events=TRUE, smlab=" ", xlab="Odds
Ratio", ref=1, overall=TRUE, print.I2=T,  leftlabs=c("Study", "Cancer cases",
"Controls", "OR", "95% CI"), xlim=c(0.25, 4), print.byvar=F, col.square="black",
col.by="black", addspace=TRUE, print.tau2=FALSE, rightcols=FALSE,
leftcols=c("studlab", "n.e", "n.c", "effect", "ci"), text.random="Pooled estimate",
ff.random.labels=2, sortvar=me$studlab, ff.random=1, ff.hetstat=1, fontsize=11,
squaresize=1)
dev.off()

# Subset: > 20
me<-metagen(TE=Estimate[cigarette_day=="> 20"], seTE=StdErr[cigarette_day==">
20"], studlab=study[cigarette_day=="> 20"],  sm="OR",
n.e=CASES[cigarette_day=="> 20"], n.c=CONTROLS[cigarette_day=="> 20"],
byvar=area[cigarette_day=="> 20"], level=0.95, level.comb=0.95,
comb.random=TRUE, comb.fixed=FALSE, method.tau="DL", title="> 20")

setwd("F:\\DOTTORATO\\RELAZIONE TERZO ANNO\\RESULTS\\Smoking
forest plot pdf\\")
pdf(file="cigarette_day 4- up 20.pdf", paper = "a4r",  width = 10,  height = 10,
pagecentre=T,  pointsize=10)
forest.meta(me, pooled.totals=TRUE, pooled.events=TRUE, smlab=" ", xlab="Odds
Ratio", ref=1, overall=TRUE, print.I2=T, leftlabs=c("Study", "Cancer cases",
"Controls", "OR", "95% CI"), xlim=c(0.25, 4), col.square="black", col.by="black",
addspace=TRUE, print.tau2=FALSE, rightcols=FALSE, leftcols=c("studlab", "n.e",
"n.c", "effect", "ci"), text.random="Pooled estimate", ff.random.labels=2,
ff.random=1,ff.hetstat=1,fontsize=11,squaresize=1)
dev.off()

tiff(filename = "F:\\DOTTORATO\\RELAZIONE TERZO
ANNO\\RESULTS\\Smoking forest plot tiff\\cigarette_day 4- up 20AREA.tiff",
width = 230, height = 230, units = "mm", res=300)
forest.meta(me,pooled.totals=TRUE,pooled.events=TRUE,smlab=" ",xlab="Odds
Ratio",ref=1,overall=TRUE,print.I2=T, leftlabs=c("Study","Cancer
cases","Controls","OR","95% CI"),xlim=c(0.25,4),print.byvar=F,
col.square="black",col.by="black",addspace=TRUE,print.tau2=FALSE,rightcols=FA
LSE, leftcols=c("studlab","n.e","n.c","effect","ci"),text.random="Pooled
estimate",ff.random.labels=2,sortvar=me$studlab,ff.random=1,ff.hetstat=1,fontsize=1
1,squaresize=1)
dev.off()
```