**Universidade de São Paulo
Faculdade de Saúde Pública**

**Equações estruturais aplicadas a modelos causais de
câncer de pulmão**

Valéria Troncoso Baltar

**Tese apresentada ao Programa de Pós-
Graduação em Saúde Pública para
obtenção de título de Doutor em Saúde
Pública**

Área de concentração: Epidemiologia

**Orientador: Prof. Dr. Júlio César
Rodrigues Pereira**

**São Paulo
2011**

Equações estruturais aplicadas a modelos causais de câncer de pulmão

Valéria Troncoso Baltar

**Tese apresentada ao Programa de Pós-
Graduação em Saúde Pública da
Faculdade de Saúde Pública da
Universidade de São Paulo para
obtenção de título de Doutor em Saúde
Pública**

Área de concentração: Epidemiologia

**Orientador: Prof. Dr. Júlio César
Rodrigues Pereira**

**São Paulo
2011**

É expressamente proibida a comercialização deste documento, tanto na sua forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano da tese.

Para meus pais, Eugenia e Paulo.

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer aos professores Júlio Pereira e Paolo Vineis pela orientação. Ao Júlio, orientador e amigo de uma vida (meu chefinho) que participou em todas as decisões profissionais que eu tomei desde que nos conhecemos. Ao Paolo, pela orientação e apoio nesse ano de aprendizado na Imperial College. Agradeço ainda aos professores Dirce Marchioni, Lúcia Barroso, Gizelton Alencar e Sergio Koifman, pelas críticas e sugestões na pré-banca, que ajudaram a melhorar este trabalho.

Também não posso deixar de registrar meus agradecimentos a Annelie e a Cristina pela ajuda na revisão da versão final.

Ao pessoal do LEE, em especial, à Helen, ao Valdir e à Robertinha pela amizade de sempre. Às minhas amigas, Rejane e Cris, sempre me incentivando.

Ao pessoal da salinha: Clóvis, Angelinha, Dani, Rossana, Mônica, Raquel e Tati, pelo companheirismo; e em especial à Bruna, pelo apoio integral na reta final. Ao pessoal das secretarias, principalmente à Angela, à Renilda e ao Fernando.

Aos amigos da Imperial College, especialmente à Wei, Rosa, Rui, Aneire, Doris, Anne-Claire, Shu-Chun e Karin, pela amizade e pelas discussões que enriqueceram este trabalho.

Aos meus amigos que compreenderam minhas ausências e me incentivaram nos momentos mais difíceis. A todos os amigos de Londres, que também foram importantes pra conclusão desta tese. A todos que, direta ou indiretamente, colaboraram nesses 4 anos de doutorado.

À toda minha família, em especial aos meus pais que me deram total apoio, financeiro e moral, quando decidi me dedicar exclusivamente ao doutorado. Aos meus irmãos, Carol e Paulão, pela cumplicidade de sempre. Às minhas primas-irmãs, Mari e Liginha, sempre atenciosas e me dando muito carinho. Gostaria de lembrar a querida tia Bel, com saudades, pelas longas conversas que me ajudaram a decidir o que fazer da minha vida profissional. À toda minha família, do Recife e do Chile.

Por fim, agradeço ao CNPq e à CAPES pelo auxílio financeiro

RESUMO

Baltar VT. **Equações estruturais aplicadas a modelos causais de câncer de pulmão.** São Paulo; 2011. [Tese de doutorado – Faculdade de Saúde Pública da USP]

Introdução: O câncer de pulmão (CP) é o tipo de câncer que mais mata no mundo e o cigarro ainda é sua causa mais importante. Além disso, a alimentação tem sido associada ao CP, por ser fonte de vitaminas e aminoácidos que fazem parte do metabolismo do carbono (MC). O MC é considerado mecanismo chave na manutenção da integridade do DNA e na regulação da expressão gênica, que, dessa forma, deve estar relacionado à carcinogênese. A ativação da imunidade está associada ao envelhecimento em indivíduos saudáveis, assim como a uma série de patologias, incluindo o câncer.

Objetivo: Estudar como o MC, a ativação da imunidade e o tabaco estão relacionados ao risco de CP em um estudo caso-controle aninhado à coorte do EPIC (*European Prospective Investigation into Cancer and Nutrition*).

Métodos: Para avaliar se os níveis plasmáticos de cotinina são um bom biomarcador da exposição ao tabaco, foram utilizados modelos lineares generalizados. Para avaliar os efeitos do tabaco, do MC e da ativação da imunidade no risco de CP, foram aplicados modelos de equações estruturais (MEE) de duas maneiras diferentes (com e sem variáveis latentes).

Resultados: Com base nas respostas aos questionários de qualidade de vida, com relação às questões sobre fumo ativo e passivo, a cotinina se mostrou um bom biomarcador de exposição recente ao tabaco (tanto o aumento da exposição passiva quanto ativa foram significativas, $P < 0,001$ e $P < 0,001$ respectivamente). Em um MEE com variáveis observadas, incluindo o MC e a via de ativação da imunidade, a metionina e o folato como causas proximais apresentaram uma forte e inversa associação com o risco de CP. O aumento em um desvio-padrão nos níveis séricos de metionina e de folato significou uma redução no risco de 19% ($P < 0,01$) e 12% ($P = 0,03$) respectivamente. Em um MEE com variáveis latentes (cada uma representando o conjunto de vitaminas e aminoácidos importantes para promover: metilação de DNA, síntese de nucleotídeos e imune ativação), foram encontrados efeitos protetores diretos da metilação do DNA ($P = 0,018$) e da imune ativação ($P = 0,037$); por outro lado, a síntese de nucleotídeos não apresentou efeito no risco do

câncer ($P=0,098$). Nas duas abordagens de MEE o cigarro permaneceu como a causa de maior impacto.

Conclusões: A cotinina mostrou-se um bom biomarcador da exposição ao tabaco (ativa e passivamente). Confirmou-se que a via de metilação é um fator de proteção contra o CP. A ativação da imunidade apresentou um efeito direto de proteção contra o CP no modelo com variáveis latentes, enquanto que, a síntese de nucleotídeos não apresentou relação com o CP. O tabaco continua sendo o fator de maior impacto no risco de CP.

Descritores: câncer de pulmão, metilação, síntese de nucleotídeos, ativação da imunidade, inflamação, modelos de equações estruturais, modelo latente.

ABSTRACT

Baltar VT. **Structural equation models applied to lung cancer causal models.** São Paulo (BR); 2011. [PhD thesis – Faculdade de Saúde Pública da Universidade de São Paulo, Brazil]

Background: Lung cancer (LC) continues to be the most common cancer death in the world. Tobacco exposure continues to be the most important cause. In addition, micronutrient intake has been linked to LC, because they are the main source of vitamins and amino acids involved in the one-carbon metabolism (OCM) which is considered key in maintaining DNA integrity, regulating gene expression, and may thus affect carcinogenesis. Immune activation is involved in the aging process in normal healthy individuals as well as in a number of pathologies, including cancer.

Objectives: To investigate how OCM, immune activation and tobacco are related to LC incidence in a nested case-control study from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort.

Methods: To validate plasma cotinine levels as a good biomarker for tobacco exposure, a generalized linear model was applied. To evaluate the effects of tobacco, OCM and immune activation in LC, structural equation models (SEM) were applied in two different ways.

Results: Based on questions about smoking, passive smoking and number of cigarettes smoked, it was shown that cotinine is a good biomarker for tobacco exposure (passive and active exposure with significant relation, $p < 0.001$ and $P < 0.001$, respectively). In a SEM model with only observed variables, including OCM and immune activation, methionine and folate as proximal causes presented a strong and inverse relation with LC risk. An increase in one standard deviation of serum levels of methionine and folate meant a 19% ($P < 0.01$) and 12% ($P < 0.01$) reduction in LC risk, respectively. In a SEM including latent variables (each one including vitamins and amino acids important to promote DNA methylation, nucleotide synthesis and immune activity), a direct and protective effect for DNA methylation ($p = 0.018$) and immune activation was found ($p = 0.037$), whereas nucleotide synthesis did not present a significant total effect. In both approaches of SEM, tobacco exposure remains with the highest impact on LC risk.

Conclusions: It was found that cotinine is a good biomarker of tobacco exposure (active and passive). It was confirmed that methylation protects against LC. Immune activation presented a direct protective effect in the latent model, while nucleotide synthesis was not confirmed to be related to LC risk. Tobacco effect remains as the factor with highest impact in lung cancer.

Keywords: lung cancer, methylation, nucleotide synthesis, immune activation, inflammation, structural equation models, latent model.

APRESENTAÇÃO

Esta tese está organizada nos seguintes capítulos: 1) Introdução; 2) Objetivos; 3) Metodologia; 4) Resultados e discussão; 5) Considerações finais; 6) Referências; e Anexo. O capítulo de resultados e discussão está apresentado em forma de artigos científicos. São três artigos, sendo que os dois primeiros já foram submetidos para publicação em revistas científicas. Após os resultados e discussão dos artigos, há um capítulo de considerações finais. As referências e o anexo apresentam-se no final da tese (porém, as referências dos manuscritos encontram-se nos mesmos).

ÍNDICE

1. INTRODUÇÃO	13
1.1. MECANISMOS CAUSAIS EM CÂNCER DE PULMÃO	16
1.1.1. O Tabaco	17
1.1.2. A Metilação de DNA	18
1.1.3. A Síntese de Nucleotídeos	20
1.1.4. A Ativação da Imunidade	20
1.1.5. A Transulfuração	21
1.2. MODELOS CAUSAIS EM EPIDEMIOLOGIA	21
1.3. O MODELO DE EQUAÇÕES ESTRUTURAIS	29
2. OBJETIVOS	34
3. METODOLOGIA	36
3.1. MATERIAL	36
3.2. MÉTODOS	37
4. RESULTADOS E DISCUSSÃO	39
4.1. MANUSCRITO 1	39
4.2. MANUSCRITO 2	59
4.3. MANUSCRITO 3	86
5. CONSIDERAÇÕES FINAIS	105
6. REFERÊNCIAS	108
ANEXO A - Medidas de Sinergia	116
ANEXO B - Cartas do Comitê de Ética em Pesquisa	129
ANEXO C - Curriculum Lattes	131

Lista de Figuras

Figura 1 - Metilação de histonas	19
Figura 2 - Diagrama de caminhos com efeito direto e indireto de X em Y	29
Figura 3 - Diagrama de caminho com três variáveis latentes	31
Figura A1 - Esquematização conceitual de três conjuntos suficientes para uma doença (ROTHMAN, 1976b)	117
Figura A2 - Esquematização conceitual de quatro classes envolvendo duas causas componentes para uma doença (ROTHMAN, 1986)	118

Siglas utilizadas

EPIC – European Prospective Investigation into Cancer and Nutrition

MEE - Modelo de equações estruturais

MEM –Modelo de equações marginais

MLG - Modelos lineares generalizados

1. INTRODUÇÃO

Em 2008, no projeto GLOBOCAN (IARC 2008), estimou-se que 12,7% dos novos casos de câncer eram de pulmão. Aproximadamente 85% dos cânceres, denominados carcinomas, ocorreram em células epiteliais. Entre 90 e 95% dos cânceres de pulmão são carcinomas de pequenas células ou carcinomas de células não-pequenas. O câncer de pulmão de pequenas células, que atinge 10-15% do total das pessoas com câncer de pulmão, é considerado patologicamente diferente dos demais tipos por ser biologicamente agressivo, de natureza difusa, com propensão a metastase precoce e prognóstico desfavorável (WCRF/AICR 2007). O câncer de pulmão é um tipo de câncer mais comum em países desenvolvidos, devido ao fato de esses países terem sofrido a epidemia do cigarro mais cedo, e é um dos cânceres que mais mata, com uma sobrevida média em 5 anos de apenas 10% (WCRF/AICR 2007).

As causas do câncer de pulmão são diversas. Incluem histórico familiar de câncer, exposição ao tabaco, a agentes infecciosos, a medicação, a radiação, a químicos industriais e também a agentes carcinogênicos presentes em comidas e bebidas (WCRF/AICR 2007). Essas causas são conhecidas por influenciarem o processo carcinogênico devido a modificações na estrutura e função celular.

Cada célula do corpo humano contém o genoma completo dentro de cada gene, porém é a expressão gênica (processo que ativa ou inativa a informação de um gene) que controla quais e quando subgrupos de genes serão ativados, permitindo que as células se diferenciem quanto à estrutura e função (WCRF/AICR 2007, FERRAZ et al. 2010).

Mudanças na expressão gênica contida no DNA são importantes para evolução e adaptação dos seres vivos; contudo, algumas dessas mudanças podem resultar em câncer, que se caracteriza pelo crescimento celular descontrolado. A divisão, diferenciação e a morte celular são processos cuidadosamente regulados para garantir crescimento, desenvolvimento e funções celulares adequadas (WCRF/AICR 2007). Todo câncer tem início em uma única célula que perde o controle de seu processo de replicação e crescimento.

Os avanços no campo da biologia molecular têm ajudado a entender os mecanismos envolvidos nesses processos celulares. Vários novos campos de pesquisas se destacam, são os chamados –ômicos. A *nutrigenômica* estuda os impactos da nutrição na manutenção da estabilidade genômica e também a influência do genótipo na determinação da resposta à dieta; a *epigenômica* nutricional estuda os perfis de expressão gênica em resposta à nutrição; a *transcriptômica* nutricional estuda os perfis de expressão gênica no nível do RNA influenciados pela nutrição; a *proteômica* estuda as proteínas que podem ser expressas pelas células; e a *metabolômica* estuda os processos metabólicos na célula e a regulação metabólica em células e tecidos (WCRF/AICR 2007).

A integridade celular depende de um balanço entre a proliferação e a morte celular e também uma diferenciação celular apropriada; Dessa forma, alterações genéticas e epigenéticas são relacionadas ao câncer (FERRAZ et al. 2010). Uma alteração genética é caracterizada pela mutação gênica, na qual a sequência do código do DNA é alterada e é passada de célula mãe para filha (na multiplicação celular, o câncer prolifera). Essa mutação pode ser a consequência de um erro que não foi reparado na duplicação celular. Uma alteração epigenética é uma alteração

que não muda o código genético, mas também pode ser repassada de célula mãe para filha. Esse tipo de alteração altera a expressão gênica, ou seja, as células podem perder suas funções essenciais. Metilação do DNA, acetilação do DNA, mudanças na homeostase das histonas, micro RNA e estabilidade do DNA são exemplos de alterações epigenéticas (WCRF/AICR 2007).

Esses dois tipos de alteração são influenciados por causas ambientais. Para o câncer de pulmão, a causa ambiental mais convincente é a exposição aos químicos do tabaco. Estima-se que o cigarro (fumo voluntário ou involuntário) seja responsável por 85% dos casos de câncer de pulmão (WHO 2006). Outras causas químicas, como a exposição ao arsênio, também são bem estabelecidas. Além dessas causas, fatores ligados a qualidade de vida vêm sendo bastante estudados, tais como atividade física e fatores nutricionais (WCRF/AICR 2007, MARCHIONI et al. 2008). STEFANI et al. (2008) concluíram, em um estudo caso-controle de homens uruguaios, que um padrão nutricional rico em anti-oxidantes (que incluiu também marcadores de frutas e vegetais) protege contra o câncer de pulmão entre ex-fumantes (essa evidência não foi encontrada para fumantes, sugerindo que fumantes extremos confundiram esse efeito) e, por outro lado, um padrão nutricional rico em carnes indicou maior risco, sendo que esses padrões foram elaborados com base em questionário de frequência alimentar e análise fatorial. Outro fator importante é a infecção por vírus, parasitas e bactérias, que podem estar ligados ao câncer, devido a danos no DNA.

Este estudo objetiva avaliar os efeitos das vitaminas e aminoácidos ligados a uma série de reações químicas que acabam por afetar fatores genéticos e epigenéticos ligados ao câncer. Para tanto, foi analisada uma base de dados de um estudo caso-

controle aninhado à coorte do “*European Prospective Investigation into Cancer and Nutrition*” (EPIC). A importância da identificação de causas de alterações epigenéticas é que estas podem ser reversíveis. Além do efeito do cigarro, que é a causa de maior risco atribuível e que é incluído no estudo medindo-se o efeito da cotinina (biomarcador de exposição à nicotina – manuscrito 1), dois novos modelos de estimação do risco de câncer de pulmão são propostos em dois artigos separadamente (manuscritos 2 e 3).

Este trabalho justifica-se pela importância da identificação de causas para uma doença tão prevalente e com elevada mortalidade como é o câncer de pulmão. A inclusão do tabagismo, reconhecido pela comunidade científica como uma importante causa do câncer de pulmão, é também relevante devido à necessidade de políticas de proibição do fumo em áreas públicas, em razão de seus efeitos em fumantes passivos (manuscrito 1). A identificação de causas relacionadas à nutrição é de grande valia para nortear políticas públicas de orientação alimentar. Ressalta-se que este estudo, além de propor MEE para confirmação de causas do câncer de pulmão, mostra a importância de técnicas que abranjam não somente as causas diretas, mas importantes cofatores causais que ajudam no entendimento do mecanismo causal envolvido no processo de saúde-doença (manuscritos 2 e 3).

1.1 MECANISMOS CAUSAIS EM CÂNCER DE PULMÃO

Consideram-se quatro mecanismos que podem estar ligados ao risco de câncer de pulmão: a alteração epigética de metilação de DNA, que quando muito baixa

(globalmente) ou muito alta (em um gene específico) pode desproteger ou silenciar os genes; a síntese de nucleotídeos (renovação celular), a ativação da imunidade, que está ligada a processos infecciosos e inflamatórios, e a transulfuração, que pode influenciar nos níveis de homocisteína e na síntese do antioxidante glutatona. Além desses, o tabaco, sendo a causa de maior impacto no risco do câncer de pulmão, com um risco atribuível de 85%, não pode deixar de ser considerado.

1.1.1 O Tabaco

O cigarro contém pelo menos 80 tipos de substâncias carcinogênicas e mutagênicas, incluindo arsênio, cádmio, amônia, formadeído e benzopireno (WCRF/AICR 2007). Cada um desses químicos tem um mecanismo diferente na causa do câncer. O benzopireno, por exemplo, pode formar adutos de DNA (em inglês: “*DNA adducts*”) em células epiteliais do pulmão. Esses adutos alteram a estrutura e interrompem a multiplicação do DNA, causando mutação e até supressão de material genético.

Apesar de o risco atribuído ao cigarro ser bastante alto, apenas 10-15% dos fumantes desenvolvem o câncer de pulmão. Isso sugere que há uma variação na susceptibilidade individual para se adquirir a doença (LO et al. 2010). É possível que alguns polimorfismos tenham um efeito na capacidade de reparo do DNA, mudando a susceptibilidade para esse câncer. LO et al. (2010) encontraram uma interação entre altos níveis de exposição passiva ao tabaco e o polimorfismo do gene MLH1-93A>G, em pessoas que nunca fumaram. Os autores advertem que há uma possibilidade de que essa associação seja resultado de uma metilação aberrante com

consequente silenciamento do gene, mas que esse estudo é mais uma evidência que se soma, sugerindo que o câncer de pulmão entre aqueles que nunca fumaram apresenta uma etiologia e uma história natural diferentes. LO et al. (2010) sugerem que indivíduos com esse polimorfismo possivelmente sejam suscetíveis a contrair esse câncer após baixos níveis de exposições a fatores carcinogênicos. Em um estudo de meta-análise (JI et al. 2010), o polimorfismo APE1 Asp148Glu não apresentou associação com o risco de câncer de pulmão entre asiáticos e brancos, porém essa associação foi significativa entre fumantes.

1.1.2 A Metilação de DNA

A metilação do DNA é uma reação química que adiciona um grupo metil (-CH₃) a proteína no núcleo do DNA (a proteína mais comum é a histona). Essa adição do metil à histona é uma reação normal que deve ser mantida em equilíbrio. O metil adicionado à histona, que é envolta pela cromatina, protege o núcleo do DNA de ataques de vírus ou bactérias. Por outro lado, quando a metilação é exagerada, condensa a cromatina, deixando a histona inacessível, e isso pode prejudicar a expressão gênica (Figura 1). A principal consequência da hiper-metilação do DNA é que, se regiões promotoras são atingidas, genes supressores de tumores podem ficar inativos (silenciados). A hiper-metilação pode também prejudicar a capacidade de reparação do DNA (DAS E SIGNAL 2004, STOVER 2009, ULRICH et al. 2008).

A metilação do DNA tem, como doador universal de metil, o cofator enzimático S-adenosilmetionina, que é formado a partir do aminoácido metionina. A principal fonte natural de metionina são as proteínas das carnes, castanhas e sementes de

hortaliças, que devem ser repostas diariamente. Outra forma de se obter metionina é pela re-metilação de homocisteína, produzida no corpo, que recebe o grupo metil do folato (vitamina B9), na forma 5-metil tetrahydrofolato, ou da Colina (Betaína). A doação do metil pelo folato depende da presença de vitamina B12. Esse processo é conhecido como ciclo do carbono, que envolve dois ciclos: ciclo do folato e ciclo da metionina-homocisteína. A Figura 1 ilustra a transcrição da cromatina, que tende a ser hiper-metilada ou hipo-metilada. Se o gene for silenciado, esse silenciamento (inativação do gene) será passado para as células filhas (DAS e SINGAL 2004).

Figura 1. Metilação de histonas (RL-UFL 2011)

1.1.3 A Síntese de Nucleotídeos

A síntese de nucleotídeos, bases nitrogenadas, de purinas (adenina e guanina encontradas em DNA e RNA) e pirimidinas (uracila encontrada em RNA e tRNA, citosina em DNA e RNA e timina em DNA e tRNA) são importantes porque são fonte de energia para diversas reações. O ácido fólico, quando ingerido e transformado em tetrahidrofolato, recebe um grupo metil da serina (que se transforma em glicina) e fica na forma 5,10-metil-tetrahidrofolato. Essa forma de folato é essencial para converter uracila em timidina (dUMP em dTMP), através da *enzima timidilato sintetase*. Essa síntese é importante para evitar o acúmulo de uracila no DNA, que é mutagênica (VISNES et al 2009, FENECH 2010). O excesso de uracila pode gerar sequências equivocadas do tipo U:G, em vez de U:T que, se não reparadas, geram mutação (VISNES et al 2009, FENECH 2010).

1.1.4 A Ativação da Imunidade

Vírus, bactérias e parasitas podem estar associados ao câncer, devido à possibilidade de causarem dano ao DNA. Alguns vírus podem atuar no DNA, bloqueando a ação de genes supressores tumorais, e também podem desencadear a ativação de proto-oncogeneses em oncogenes (alteração do proto-oncogenes, que têm a função de controlar a divisão celular). Além disso, a resposta à infecção é a inflamação, que, em sua fase aguda, pode até ser benéfica, mas, quando crônica, pode resultar em dano ao DNA. Um ambiente inflamado está propenso a proliferação e diferenciação celular (WCRF/AICR 2007).

1.1.5 A Transulfuração

A transulfuração também faz parte do metabolismo do carbono. A homocisteína quando não é usada na re-metilação da metionina, pode ser utilizada na síntese de cistationina. Para tanto a serina é necessária para o transporte do carbono para a homocisteína e a vitamina B6 participa como cofator pelo fato da enzima cistationina β -sintetase ser dependente dela. Posteriormente a cistationina é usada para síntese de cisteína com vitamina B6 como cofator enzimático. A importância dessa parte do metabolismo do carbono deve-se ao fato do aminoácido cisteína ser essencial para síntese do antioxidante glutathionina. Quando a cisteína não é obtida pela dieta, é importante que a via de transulfuração esteja ativa (DAVIS et al 2006).

1.2 MODELOS CAUSAIS EM EPIDEMIOLOGIA

A inferência causal em saúde é de interesse de médicos, epidemiologistas, filósofos, sociólogos, e estatísticos. Nas últimas décadas, houve um aumento da contribuição destes últimos no desenvolvimento de métodos epidemiológicos, devido, principalmente, aos avanços computacionais, que facilitam a aplicação de técnicas mais complexas de análise. Modelos lineares, modelos lineares generalizados, modelos não lineares, entre outros, são bastante usuais entre epidemiologistas que buscam associar fatores de risco às doenças. Entretanto, nas revistas científicas mais importantes em epidemiologia, os modelos causais gráficos (diagramas de caminho) (PEARL 1995, WRIGHT 1921), o modelo de conjuntos

suficientes de causas componentes de ROTHMAN (1976a, 1976b, 2002), o modelo contrafactual ou de “resposta-potencial” (no inglês: *conterfactual* ou *potencial outcome*, LITTLE e RUBIN 2000, MALDONADO e GREENLAND 2002, GREENLAND e BRUMBACK 2002) e os modelos de equações estruturais (MEE) (JÖRESKOG 2002, BOLLEN 1993, MUTHÉN e SATORRA 1995, SKRONDAL e RAB-HESKETCH 2004) têm se destacado na discussão sobre causalidade.

Em 1918, Sewall Wright propôs o método de coeficientes do diagrama de caminhos, publicado de forma mais completa em 1921, no livro “Correlação e Causação” (do inglês: *Correlation and Causation*) (WRIGHT 1921). Wright montou um sistema biológico completo, no qual “todas” as causas relevantes foram incluídas. O modelo serve não só para confirmar os fatores causais hipotetizados, como também para medir seus efeitos, e, segundo o autor, não deve ser usado para descobrir causas. Wright foi ignorado por seus contemporâneos e fortemente criticado por Ronald A. Fisher (citado por YU 2007) e pelo filósofo Henry E. Niles (NILES 1922), que diziam que o modelo de caminhos assume uma estrutura causal *a priori*, mas não propõe um mecanismo de identificação de fatores causais relevantes e, por isso, não poderia ser considerado um modelo causal verdadeiro.

WRIGHT (1921) argumentou que até então a correlação de Pearson era uma estatística simétrica, uma covariação não direcionada: $A \leftrightarrow B$, e que em modelos causais isso deveria ser mudado. Com base em variáveis biológicas cujas relações sabia serem sequências irreversíveis, ele deduziu uma direção causal: $A \rightarrow B$ (A causa B). O coeficiente de caminho proposto por ele era simplesmente o coeficiente de regressão padronizado (coeficiente de correlação parcial). Sua publicação formalizou o diagrama de caminhos com um conjunto de equações simultâneas.

Com a descrença no método proposto por Wright, tanto pela escola de Pearson, que buscava a descrição das relações sem se preocupar com as causas, como pela de Fisher, que, nessa época, estava fortemente interessada em estudos experimentais aleatorizados, não apreciando um método que se baseava “apenas” no coeficiente de correlação de Pearson, os diagramas de caminhos ficaram esquecidos (YU 2007).

O redescobrimento do modelo de caminhos deu-se em 1964, quando o sociólogo Hubert Blalock (citado por YU 2007) usou correlações parciais para analisar relações causais em estudos sociológicos. Em 1970, na “Conferência de sociólogos em Madison” em Wisconsin, Joreskog apresentou a ideia de *Equação estrutural linear*, que acabou se materializando no software LISREL, o primeiro software para análise de MEE.

Em 1979, David A. Kenny publicou o livro de MEE intitulado “Correlação e Causalidade” (em inglês: “Correlation and Causality”), no qual não especificou nenhuma técnica de descoberta de causas; argumentou que, antes de se testar o modelo, a estrutura causal deve estar pré-especificada e com referência às teorias *a priori*. “MEE requer uma mistura entre matemática e teoria. Apesar de haver questões matemáticas muito interessantes nesses modelos, as questões mais difíceis são aquelas que traduzem a teoria em equações. A teoria especifica a forma das equações” (KENNY 1979). Dessa forma, um modelo causal tem como etapa mais importante a formulação do modelo teórico *a priori* e, a partir daí, os efeitos podem ser medidos quantitativamente.

A partir da década de 1980, os MEEs passam a ser utilizados como modelos causais, cujas equações são assimétricas (JAMES et al. 1982). A equação $Y=a+bX$ representa uma função de causa-efeito, onde X's são as causas e Y o efeito. A teoria

tem um papel fundamental na formulação da ordem causal e pode ser usada para identificar variáveis relevantes em um sistema completo no qual todas as causas são incluídas. Os MEEs testam, confirmam uma teoria causal, mas não a descobrem. E o modelo é dito confirmado quando é consistente com os dados (no caso dos MEEs, ser consistente significa que a matriz de correlação observada não difere muito da proposta pelo modelo) (MELHADO 2009).

Os MEEs utilizam os diagramas de caminho como um auxílio para formulação teórica dos modelos. Em epidemiologia, os modelos gráficos de caminho foram popularizados por GREENLAND e BRUMBACK (2002), GREENLAND et al. (1999), PEARL (1995, 2000) e ROBINS (2001). Usualmente, esses gráficos são representados por letras e setas. As pontas das setas ligam-se às variáveis, representadas por letras, ditas *efeitos*; e as caudas conectam-se as variáveis chamadas de *causas*.

Esses diagramas são extremamente úteis para visualização de relações causais, que podem ser diretas, ou podem apresentar um efeito mediador. O efeito mediador, Z, é efeito de uma causa anterior e é causa de um efeito seguinte: $X \rightarrow Z \rightarrow Y$. Quando os caminhos não se fecham, ou seja, não é formado um ciclo, não há um caminho que comece e volte para uma mesma variável, e, se as sequências são diretas, ou seja, se o efeito de uma sequência é a causa da sequência seguinte, formam-se os chamados *gráficos acíclicos diretos* (em inglês, DAGs: “direct acyclic graphs”). Não é necessário que se incluam todas as causas no diagrama, mas, se duas ou mais variáveis do gráfico têm a mesma causa, essa causa deve estar no gráfico como uma causa de ambas, se não o gráfico pode ser considerado não causal (GREENLAND e BRUMBACK 2002).

Os diagramas de caminhos são uma forma simples e de fácil compreensão para representar as direções causais e também a independência entre as variáveis. Porém, havia necessidade de uma maior formalização quantitativa, com suposições mais fortes desses modelos gráficos, que eram até então mais filosóficos do que estatísticos. No início do século XX, Neyman e Fisher propuseram o modelo contrafactual ou de resposta-potencial (citado por LITTLE e RUBIN 2000). Segundo MACMAHON e PUGH (1967, pg 12), “...uma associação pode ser classificada presumptivamente como causal quando acredita-se que, se a causa (exposição) for alterada, o efeito teria mudado”. Essa frase descreve o elemento contrafactual, que seria o que teria acontecido se, ao contrário do fato, a exposição tivesse sido outra diferente do que ela realmente foi (GREENLAND E BRUMBACK 2002). A resposta potencial para um indivíduo i e tratamento (exposição) x_j é a resposta que o indivíduo teria no momento do tratamento, dado que esse tratamento existe, mesmo que o indivíduo de fato não tenha recebido esse tratamento (resposta potencial do indivíduo i sob tratamento x_j) (GREENLAND e BRUMBACK 2002). Essa pressuposição é controversa, devido ao fato de que só é possível observar essa resposta naqueles que de fato receberam o tratamento (as demais respostas, para os que não receberam tratamento, podem ser apenas estimadas, e não observadas). Essa necessidade de premissas que não podem ser testadas é desconfortável, mas os especialistas em modelos *contrafactuais* se defendem, dizendo que é um problema geral de inferência causal, sendo que, em modelos contrafactuais, esse problema é colocado de forma explícita (GREENLAND E BRUMBACK 2002, MALDONADO e GREENLAND 2002, GREENLAND et al. 1999). Para que esse tipo de modelagem possa ser utilizado na prática, faz-se uso da teoria da probabilidade, estimando-se, a

partir de uma amostra aleatória, a distribuição da resposta em uma população em diferentes níveis do tratamento ou da exposição, da mesma maneira como é feito em modelos estatísticos usuais (GREENLAND e BRUMBACK 2002).

Com o início de estudos das doenças degenerativas e com o abandono de estudos epidemiológicos unicausais, o conceito de interação passou a ocupar uma posição bastante central nas discussões de causalidade. Três tipos de interação se destacaram: i) interação estatística; ii) interação por excesso de risco; e iii) interação biológica (GREENLAND e BRUMBACK 2002). i) a interação estatística estará sujeita ao tipo de modelo estatístico utilizado, ela pode ser um efeito multiplicador ou aditivo, dependendo da função de ligação adotada no modelo linear; ii) a interação por excesso de risco pode ser diretamente acessada apenas em modelos que não façam uso de transformação logarítmica; assim, o excesso de risco é aditivo e não multiplicativo. Em um modelo de regressão logística, por exemplo, o acesso à interação por excesso de risco pode ser feito com o auxílio de estatísticas de sinergia (ver Anexo); E por fim, iii) a interação biológica, que é a mais complicada delas, visto que a medida do excesso de risco (sinergia) pode ajudar a identificá-la, mas uma interação biológica é um mecanismo que ocorre no nível individual, mesmo que ocorra uma anulação dos riscos quando a medida é multiplicativa ou de outra natureza, nada garante que a interação biológica ocorre para alguns indivíduos (GREENLAND 2009). O modelo de Rothman, que considera conjuntos suficientes de causas componentes é o modelo, com base em mecanismo causal, mais famoso em Epidemiologia. Rothman usa a ideia de excesso de risco em modelos aditivos para acessar a interação biológica, mas há autores que julgam esse termo “biológico” equivocado, e preferem chamar a sinergia de excesso de risco devido à interação

(GREENLAND 2009). Esse modelo tem como base o mecanismo de interação, a sinergia entre causas (ROTHMAN 1974, 1976a), que são necessárias juntas para que o mecanismo opere e produza a doença em questão (GREENLAND e BRUMBACK 2002).

Essa definição do modelo, com base em mecanismos causais, é elaborada no nível individual, similarmente aos modelos de respostas potenciais. Então, diferentes grupos de mecanismos determinam respostas potenciais para os indivíduos. Na prática, o modelo de conjuntos suficientes de causas componentes não será totalmente determinístico, visto que cada causa componente será assumida como sendo variável aleatória (POOLE 2001).

A formalização do modelo causal de Rothman deu-se em 1976, quando ele publicou a primeira proposta de estimação da sinergia (ROTHMAN 1976a) e também a formalização do modelo nos gráficos de pizza (ROTHMAN 1976b) no artigo ‘Causas’ (‘Causes’). Toda a filosofia desse modelo teórico de conjuntos suficientes é baseada na teoria desenvolvida pelo filósofo australiano John Mackie, intitulada condição INUS, das iniciais em inglês: “Cause is an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result” (MACKIE 1965, 1974).

No seu modelo de conjuntos suficientes de causas componentes ROTHMAN (1976b), define um conjunto de causas *suficientes* como um grupo de condições ou acontecimentos, não supérfluos, que produzem a doença. Completar um conjunto suficiente desencadeia a doença. Em biologia, muitas vezes, esses conjuntos nem são conhecidos. Há de se pensar, então, que existe uma interação biológica entre causas, que tem um efeito em indivíduos suscetíveis a essas causas, e que,

consequentemente, desenvolverão a doença. Como muitas causas não são conhecidas, o melhor que se pode fazer é avaliar o valor médio do risco para os expostos a um padrão dado de indicadores de risco conhecidos. Essas causas que formam o conjunto chamado de *suficiente* são ditas causas *componentes*. Algumas das causas componentes podem pertencer a diferentes conjuntos suficientes. Algumas causas componentes podem estar presentes em todos os conjuntos suficientes alternativos, e, quando isso ocorre, Rothman chama essas causas de *necessárias*.

O mecanismo que Rothman propõe, ressaltando a importância de se avaliar o efeito de duas causas conjuntas e indicando que ambas, concomitantemente, apresentam um risco diferente do que seria observado com as causas em separado nem sempre é relevante. No caso do câncer de pulmão, há um crescente interesse nos efeitos de fatores ligados à nutrição, que, depois dos efeitos de misturas químicas, como os efeitos do tabaco, vêm mostrando uma forte relação com a incidência da doença (FERRAZ et al. 2010; BALUZ et al. 2002). Essas reações não ocorrem conjuntamente, mas são uma sequência bem estabelecida de reações químicas (estudos laboratoriais ajudam a entender essas reações). Nesse caso, a importância dos diagramas de caminho, juntamente com a estimação dos efeitos causais, parece mais importante do que o modelo de causas componentes suficientes de Rothman. O MEE é o método mais apropriado para estimação dos efeitos dessas cadeias de reações estabelecidas *a priori*, com a vantagem de poder incluir as variáveis em sua forma contínua, sem a necessidade de categorização para avaliação de risco, como é necessário para as estatísticas de sinergia.

1.3 O MODELO DE EQUAÇÕES ESTRUTURAIS

O MEE, de uma maneira geral, usa os diagramas de caminho como uma forma mais simples de entender as relações causais que são baseadas em teorias. A diferença de um modelo de regressão tradicional para um MEE é que, neste, os fatores de risco podem afetar a variável resposta direta e/ou indiretamente. O modelo de caminho mais simples é $X \rightarrow Y$, que equivale a uma regressão do tipo $Y = \alpha + \beta X + \varepsilon$. Um caminho incluindo um efeito indireto de X via Z (variável mediadora ou intermediária) está representado na Figura 2. Esse modelo precisa de duas equações para ser representado matematicamente, visto que há duas variáveis resposta (o número de equações coincide com o número de variáveis resposta). A primeira regressão mede os efeitos de X em Z: $Z = \alpha_0 + \alpha_1 X + \varepsilon$. A segunda mede o efeito direto de X em Y e o efeito de Z em Y: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$.

Figura 2. Diagrama de caminhos com efeito direto e indireto de X em Y

As duas regressões representam um único modelo causal, por isso devem ser estimadas simultaneamente. Assumindo que os erros são estatisticamente independentes, essas duas equações podem ser estimadas por: $Y = \beta_0 + \beta_1 X + \beta_2(\alpha_0 + \alpha_1 X) + \varepsilon = (\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1)X + \varepsilon$. O modelo representado na Figura 2 pode, então, ser estimado com base apenas na variável X (equação na forma

reduzida) (RETFERFORD e CHOE 1993). O efeito β_1 é chamado de efeito direto de X em Y (caminho simples ou direto); o produto $\beta_2\alpha_1$ é o efeito indireto de X em Y (caminho composto) e a soma $\beta_1+\beta_2\alpha_1$ é o efeito total de X em Y.

A inclusão de variáveis de controle nas equações de regressão que compõem o sistema de equações consideradas para MEE pode ser feita de regressão em regressão (para cada variável resposta há a possibilidade de controle por algum fator), sendo que não são ilustradas nos diagramas de caminho, pois não são considerados fatores causais.

Esse tipo de modelo, utilizando apenas variáveis observáveis, é conhecido como *modelo estrutural*. Um MEE pode apresentar, também, uma abordagem não estrutural, conhecida como *modelo de mensuração*, que utiliza variáveis observáveis para estimar variáveis que não são diretamente observáveis, as *variáveis latentes* (ALENCAR 2009, MELHADO 2009). Um MEE completo apresenta essas duas partes: estrutural e de mensuração.

Modelos de variáveis latentes são muito usuais em psicologia, pois usam constructos para compor indicadores que não podem ser observados diretamente. Esse tipo de modelagem é bastante semelhante a análise fatorial, mas é uma análise confirmatória, pois as correlações entre os itens que irão compor as variáveis latentes são estipuladas *a priori* (LONG 1983). O MEE testará se essas suposições se confirmam.

Quando um MEE inclui essas duas partes da modelagem, geralmente as variáveis latentes são representadas por círculos nos diagramas de caminho, e as variáveis observadas são representadas por retângulos. Na Figura 3, está representado um MEE com três variáveis latentes: F1, F2 e F3. F1 é composta por três variáveis

observáveis: A, B e C. F2 é composta por C, D e E. F3 é composta por C, E e F. Nesse modelo, não é hipotetizado que as variáveis latentes são correlacionadas entre si (flechas duplas representariam essa correlação: $\leftrightarrow$). Os erros Er3 e Er4 apresentam correlação. O evento final é a ocorrência da doença, que é uma variável observada, e apresenta um erro de mensuração: Er.

Figura 3. Diagrama de caminho com três variáveis latentes

Esses modelos foram desenvolvidos inicialmente para variáveis observáveis quantitativas e depois também para variáveis categorizadas. Em saúde, é comum tratar-se da ocorrência ou não de um evento, e a combinação de MEE com modelos lineares generalizados é bastante prática. Essa parte da discussão refere-se à estimação do modelo, mas antes disso é necessário um comentário sobre identificabilidade. A identificabilidade do modelo é o primeiro passo que deve ser levado em consideração e consiste em saber se a solução alcançada é única ou não. Um modelo não identificável pode não ter solução ou ter muitas (pode apresentar,

por exemplo, infinitos valores para seus parâmetros), mas o modelo identificável terá apenas uma solução. Quando ocorre uma estimação que é não identificável, esse modelo não faz sentido, a interpretação de seus parâmetros pode ser absurda. Algumas restrições podem ser impostas ao modelo para garantir a sua identificabilidade. Sendo o modelo identificável, parte-se para estimação (ALENCAR 2009, MELHADO 2009). No software Mplus, escolhido para análise dos dados desta tese, a estimação de MEE permite o uso de variáveis observáveis do tipo contínuo, binário, ordinal, nominal e de contagem. As variáveis latentes podem ser do tipo contínuo ou categórico. Para as variáveis resposta contínuas, são utilizados modelos de regressão linear. Para variáveis respostas binárias ou categóricas ordinais, são utilizadas regressões logísticas e probito. No caso de variáveis respostas nominais, usa-se regressão logística multinomial e, para respostas do tipo contagem, regressão de Poisson ou binomial negativa (com ou sem zero inflacionado).

Como os MEEs baseiam-se na estrutura de correlação observada, a definição correta do tipo de variável é importante para que as correlações apropriadas sejam utilizadas na matriz de correlação observada. No Mplus, a matriz de correlação usa a correlação apropriada para cada par de variáveis analisado. A correlação de Pearson é usada entre duas variáveis contínuas; a correlação “tetrachonic” (HARRIS, 1988) é utilizada entre duas variáveis binárias; a correlação “polychoric” (JÖRESKOG, 2002) é usada entre uma variável binária e uma ordinal; a correlação “polyserial” (DRASGOW, 1988) é usada entre uma variável contínua e outra ordinal; e a correlação “biserial” (KRAEMER, 1982) quando uma variável é contínua e a outra binária.

Neste estudo, optou-se pela utilização de modelos lineares generalizados, especificamente para o uso da distribuição binomial (foram utilizadas as funções de ligação logito e também a probito, que facilitam a interpretação para resposta binária) para estimação dos efeitos proximais na ocorrência do evento.

2. OBJETIVOS

O principal objetivo deste estudo é avaliar a plausibilidade de causas nutricionais, além do efeito do tabaco, no risco de câncer de pulmão em um estudo caso-controle aninhado à coorte do “*European Prospective Investigation into Cancer and Nutrition*” (EPIC). Para tanto, aplica-se MEE visando a estabelecer esse método como um bom instrumento para confirmação de teorias causais.

Com esse objetivo, separadamente para cada um dos manuscritos apresentados nos resultados deste estudo, tem-se:

Manuscrito 1: justificar o uso de um biomarcador, a cotinina, como uma boa opção de medida de exposição à nicotina (ativa e passivamente). Verificar se o mesmo está relacionado às categorias de tabagismo atual (nunca fumou, ex-fumante e fumante), ao número médio de cigarros fumados por dia (para fumantes) e ao número de horas de exposição passiva (para não fumantes), controlando-se potenciais confundimentos.

Manuscrito 2: avaliar os efeitos de substâncias presentes no sangue humano, devido a fatores nutricionais, no risco de câncer de pulmão. Essas substâncias são relacionadas ao metabolismo do carbono (ciclo do folato e ciclo da homocisteína-metionina), à transulfuração e à ativação da imunidade, além do uso da cotinina como biomarcador de exposição a nicotina. Neste manuscrito a abordagem metodológica visa confirmar epidemiologicamente os caminhos causais biológicos estabelecidos na literatura.

Manuscrito 3: tem por objetivo a avaliação dos efeitos de substâncias, também colhidas em amostras de sangue antes do aparecimento da doença, no risco de câncer

de pulmão. A abordagem metodológica difere do manuscrito anterior pelo fato deste manuscrito objetivar a estimação de variáveis latentes que representam conjuntos de variáveis teoricamente relacionadas à: metilação do DNA, síntese de nucleotídeos e imune ativação, além da cotinina que entra no modelo como biomarcador de exposição ao tabaco.

3. METODOLOGIA

3.1. MATERIAL

O estudo EPIC é uma coorte prospectiva que envolve 10 países: França, Alemanha, Grécia, Itália, Holanda, Espanha, Reino Unido, Suécia, Dinamarca e Noruega. São 23 centros espalhados nesses 10 países que colaboram com a pesquisa. O recrutamento de voluntários ocorreu entre 1992 e 2000 e um questionário sobre fatores da dieta e de estilo de vida foi aplicado. Além disso, os voluntários tiveram suas medidas antropométricas anotadas e também uma amostra de sangue coletada. O estudo foi aprovado por todos os comitês de ética locais e também pela Agência Internacional para Pesquisa em Câncer (IARC). Os participantes dessa coorte são acompanhados em intervalos de tempo regulares e suas informações são atualizadas. Essa coorte tem 519.978 participantes recrutados e informações sobre novos casos de câncer e outras doenças são constantemente atualizadas, incluindo a mortalidade. Informações sobre as amostras de sangue foram coletadas para uma grande parte da coorte e mantidas congeladas para futuras análises. RIBOLI et al (2002) publicaram uma descrição completa sobre a população estudada e a coleta de dados do estudo EPIC.

Este estudo utiliza uma subamostra, que é um estudo caso-controle aninhado a coorte do EPIC. As amostras de sangue dos casos e de seus respectivos controles foram analisadas simultaneamente. As análises séricas e plasmáticas, das variáveis estudadas, foram analisadas pelo laboratório Bevitall S/A (www.bevital.no).

No total da coorte 2.206 casos de câncer de pulmão foram diagnosticados como incidentes (código C34 que considera os cânceres invasivos, da Classificação Internacional de Doenças para Oncologia, segunda edição). Indivíduos que não doaram sangue, ou que não tinham a data da coleta de sangue, ou que tiveram alguma história de outro câncer (com exceção de câncer de pele, não-melanoma), foram excluídos. Cada caso teve, no momento do diagnóstico, dois controles livres de câncer pareados segundo: país, sexo, data da coleta de sangue (± 1 mês, flexível em até 5 meses, caso não tenha controle disponível) e data de nascimento (± 1 ano, flexível em até 5 anos), para evitar confundimentos (JOHANSSON et al. 2010).

A amostra final, utilizada no presente estudo, que foi aprovado pelo comitê de ética da Faculdade de Saúde Pública da Universidade de São Paulo, tem 2638 indivíduos (891 casos e 1747 controles, sendo que 856 casos permaneceram com o pareamento 1:2 e 35 casos perderam um de seus controles por apresentarem problemas na informação em pelo uma das variáveis utilizadas ou na data de coleta do sangue).

3.2. MÉTODOS

Este estudo faz uso de três diferentes métodos estatísticos de análise dos dados. O Manuscrito 1 usa modelos lineares generalizados (MLG) para avaliar a relação entre as categorias de exposição ao tabaco (questionário de qualidade de vida) com os níveis plasmáticos de cotinina (biomarcador de nicotina). Três subamostras são analisadas: uma amostra com informação completa ($n=859$) sobre as questões relativas ao fumo passivo (exposto ou não); uma amostra de apenas não fumantes

com informação de horas de exposição ao fumo passivo (n=107); e uma amostra para apenas fumantes com informação sobre número médio de cigarros fumados por dia (n=832). Para essa análise foram considerados MLGs com distribuição gama e normal ambos com função de ligação identidade.

No manuscrito 2, foram considerados MEEs com variáveis observadas e o manuscrito 3, MEEs com variáveis latentes. Ambos os manuscritos buscam descrever diferentes caminhos na causação do câncer de pulmão. Esses dois manuscritos têm como objetivo relacionar substâncias, medidas nas amostras de sangue, que possam estar relacionadas com a incidência do câncer de pulmão devido à sua participação em reações de interesse: metilação do DNA, síntese de nucleotídeos, transulfuração e ativação da imunidade (uma descrição mais detalhada das variáveis utilizadas encontra-se nos manuscritos do capítulo de resultados e discussão).

4. RESULTADOS E DISCUSSÃO

4.1.MANUSCRITO 1

Artigo submetido no periódico *Cancer Epidemiology, Biomarkers & Prevention*.

Smoking, second-hand smoke and cotinine levels in a subset of EPIC cohort.

Valéria Troncoso Baltar^{1§}, Wei Xun², Shu-Chun Chuang², Caroline Relton³, Per Magne Ueland^{4,5}, Stein Emil Vollset^{6,7}, Øivind Midttun⁸, Mattias Johansson⁹, Nadia Slimani⁹, Mazda Jenab⁹, Françoise Clavel-Chapelon^{10,11,12}, Marie-Christine Boutron-Ruault^{10,11,12}, Guy Fagherazzi^{10,11,12}, Rudolf Kaaks¹³, Sabine Rohrmann¹⁴, Heiner Boeing¹⁵, Cornelia Weikert¹⁵, H. Bas Bueno-de-Mesquita¹⁶, Hendriek C Boshuizen¹⁶, Carla H. van Gils¹⁷, Petra HM Peeters¹⁷, Antonio Agudo¹⁸, Aurelio Barricarte^{19,23}, Carmen Navarro^{20,23}, Laudina Rodríguez²¹, José Maria Huerta Castaño^{22,23}, Nerea Larrañaga^{23,24}, Maria José Sánchez Pérez^{23,25}, Kay-Tee Khaw²⁶, Nick Wareham²⁷, Naomi E. Allen²⁸, Francesca Crowe²⁸, Valentina Gallo², Teresa Norat², Giovanna Tagliabue²⁹, Giovanna Masala³⁰, Salvatore Panico³¹, Carlota Sacerdote³², Rosario Tumino³³, Antonia Trichopoulou^{34,35}, Pagona Lagiou^{34,36}, Christina Bamia³⁴, Torgny Rasmussen³⁷, Göran Hallmans³⁸, Nina Roswall³⁹, Anne Tjønneland³⁹, Elio Riboli², Paul Brennan⁹, Paolo Vineis^{32,40}

¹Department of Epidemiology, Faculty of Public Health, University of São Paulo, Av. Dr Arnaldo, 715. CEP 01246-904. São Paulo, São Paulo, Brazil

²Department of Epidemiology & Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, St Mary's Campus, 5th floor, Norfolk Place W2 1PG, London, UK

³Institute for Ageing and Health, Newcastle University, Campus for Ageing and Vitality, Newcastle upon Tyne, NE4 5PL Newcastle, UK

⁴Section for Pharmacology, Institute of Medicine, University of Bergen, 9th floor, New Laboratory Building 5021, Bergen, Norway

⁵Haukeland University Hospital, Jonas Lies vei 65 N-5058. PO.Box 1, N-5021, Bergen, Norway

⁶Department of Public Health and Primary Health Care, University of Bergen, Postboks 7800, 5020, Bergen, Norway

⁷Division of Epidemiology, Norwegian Institute of Public Health, Kalfarvejen, 31 N-5018 Bergen, Norway

⁸Bevital AS, Laboratoriebygget, 9 etg., 5021 Bergen, Norway

- ⁹International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372, Lyon, France
- ¹⁰Institut National de la Santé et de la Recherche Médicale, INSERM U780-IFR169, 16 Avenue Paul Vaillant-Couturier, 94807 Villejuif, France
- ¹¹Institut Gustave Roussy, 39 rue Camille Desmoulins, 94805, Villejuif, France
- ¹²Paris South University, Villejuif, France
- ¹³Division of Cancer Epidemiology, German Cancer Research Centre, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
- ¹⁴Division of Clinical Epidemiology, German Cancer Research Centre, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
- ¹⁵Department of Epidemiology, German Institute of Human Nutrition, Arthur-Scheunert-Allee 114-116, 14558 Bergholz-Rehbrücke, Nuthetal, Germany
- ¹⁶National Institute of Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, The Netherlands
- ¹⁷Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Room Str. 6.131, PO Box 85500, 3508 GA Utrecht, The Netherlands
- ¹⁸Unit of Nutrition, Environment and Cancer, Catalan Institute of Oncology, Gran Via 199-203, 08907 L'Hospitalet de Llobregat, Spain
- ¹⁹Epidemiology, Prevention and Promotion Health Service, Public Health Institute of Navarra, Leyre 15, 31003 Pamplona, Navarra, Spain
- ²⁰Department of Epidemiology, Regional Council of Health and Consumer Affairs, Ronda de Levante 11, Murcia 3008, Spain
- ²¹Public Health and Participation Directorate, Health and Health Care Services Council, C/ Ciriaco Miguel Virgil no 9, CP 33006 Oviedo, Asturias, Spain
- ²²Department of Epidemiology, Murcia Regional Health Authority, Ronda de Levante 11, Murcia 3008, Spain
- ²³CIBER of Epidemiology and Public Health, C/ Dr. Aiguader, 88, 1a planta – 08003, Barcelona, Spain
- ²⁴Epidemiology Unit, Public Health Department of Gipuzkoa, Basque Government Avenida de Navarra, 4-20013 Donostia, San Sebastián, Spain
- ²⁵Cancer registry of anada, Andalusian School of Public Health, Cuesta del observatorio, 4, 18080, Granada, Spain
- ²⁶Clinical Gerontology Unit, Department of Public Health and Primary Care, University of Cambridge, Level 2, F + G Block, Box 251, Addenbrooke's Hospital Hills Road, CB2 2QQ Cambridge, UK
- ²⁷MRC Epidemiology Unit, Strangeways Research Laboratory, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK
- ²⁸Cancer Epidemiology Unit, University of Oxford, Richard Doll Building, Old Road Campus, OX3 7LF Oxford, UK
- ²⁹Environmental Epidemiology and Cancer Registry Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, via Venezian , 1, I - 20133 Milano , Italia
- ³⁰Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Centre, Scientific Institute of Tuscany, Via Cosimo il Vecchio 2, 50139 Florence, Italy
- ³¹Department of Clinical and Experimental Medicine, Federico II University of Naples, Via Pansini 5, 80131 Naples, Italy
- ³²Human Genetic Foundation (HuGeF), 52, via Nizza 10123 Torino Italy,

³³Ragusa Cancer Registry, Azienda Ospedaliera "Civile M.P. Arezzo", Via Dante N° 109, 97100 Ragusa, Italy

³⁴WHO Collaborating Center for Food and Nutrition Policies, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School 75 M. Asias Avenue, Goudi, GR-115 27, Athens, Greece

³⁵Hellenic Health Foundation, 10-12 Tetrapoleos Street, GR-115 27, Athens Greece

³⁶Department of Epidemiology, Harvard School of Public Health, Harvard University 677 Huntington Ave, Boston, MA 02115. USA

³⁷Department of Radiation Sciences, Oncology, Umeå University, SE-901 87 Umeå, Sweden

³⁸Department of Public Health and Clinical Medicine, Umeå University, SE-90187, Sweden

³⁹Institute of Cancer Epidemiology, Danish Cancer Society, Strandboulevarden 49, 2100, Copenhagen, Denmark

⁴⁰Department of Epidemiology & Public Health, MRC-HPA Centre for Environment and Health, Imperial College, Norfolk Place, W2 1PG, London, UK

[§]Corresponding author

Email addresses: VTB: vbaltar@usp.br, WX: wei.xun06@imperial.ac.uk, SCC: s-c.chuang@imperial.ac.uk, CR: c.l.relton@newcastle.ac.uk, PMU: per.ueland@ikb.uib.no, SEV: Vollset@uib.no, ØM: bjorn.midttun@farm.uib.no, MJ: johanssonm@fellows.iarc.fr, NS: slimani@iarc.fr, MJ: jenab@iarc.fr, FC-C: clavel@igr.fr, M-CB-R: boutron@igr.fr, GF: guy.fagherazzi@igr.fr, RK: r.kaaks@dkfz-heidelberg.de, SR: sabine.rohrmann@ifspm.uzh.ch, HB: boeing@mail.dife.de, CW: weikert@dife.de, HBBd-M: Bas.Bueno.De.Mesquita@rivm.nl, HCB: Hendriek.Boshuizen@rivm.nl, CHvG: C.vanGils@umcutrecht.nl, PP: P.H.M.Peeters@umcutrecht.nl, AA: a.agudo@iconcologia.net, AB: abarricg@cfnavarra.es, CN: carmen.navarro@carm.es, LR: laudina.rodriquezsuarez@asturias.org, JMHC: jmhuerta.carm@gmail.com, NL: epidem3-san@ej-gv.es, MJSP: mariajose.sanchez.easp@juntadeandalucia.es, K-TK: kk101@medschl.cam.ac.uk, NW: nick.wareham@mrc-epid.cam.ac.uk, NA: Naomi.allen@ceu.ox.ac.uk, FC: francesca.crowe@ceu.ox.ac.uk, VG: v.gallo@imperial.ac.uk, TN: t.norat@imperial.ac.uk, GT: giovanna.tagliabue@istitutotumori.mi.it, GM: g.masala@ispo.toscana.it, SP: spanico@unina.it, CS: carlotta.sacerdote@cpo.it, RT: rtumino@tin.it, AT: atrichopoulou@hhf-greece.gr, PL: pdlagiou@med.uoa.gr, CB: cbamia@nut.uoa.gr, TR: torgny.rasmuson@onkologi.umu.se, GH: goran.hallmans@nutrires.umu.se, NR: roswall@cancer.dk, AT: annet@cancer.dk, ER: e.riboli@imperial.ac.uk, PB: brennan@iarc.fr, PV: p.vineis@imperial.ac.uk

Abstract

Background

Several countries are discussing new legislation on the ban on smoking in public places, based on the growing evidence of the hazards of second-hand smoke (SHS) exposure. The objective of the present study is to quantitatively assess the relationship between smoking, SHS and plasma cotinine levels in the large European Prospective Investigation into Cancer and Nutrition (EPIC) cohort.

Methods

From a study on lung cancer in the EPIC cohort, questionnaire information on smoking was collected at enrolment and cotinine was measured in plasma. Three statistical models were applied considering samples available in a cross-section design: 1) cotinine levels by categories combining smoking and passive smoking (n=859); 2) non-smokers only to evaluate the effect of hours of passive smoking exposure (n=107); 3) current smokers to evaluate the effect of the number of cigarettes per day (n=832). All models were controlled for country, sex, age and body mass index.

Results

Amongst non-smokers, passive smokers presented significant differences in cotinine compared to non-exposed, with a bigger difference among former smokers. A one hour per day increment of passive smoking gave rise to a significant 2.58 nmol/L (0.45 ng/mL) mean cotinine increase ($P<0.001$). In current smokers, a one cigarette per day increment gave rise to a significant 22.44 nmol/L (3.95 ng/mL) increase in average serum cotinine ($P<0.001$).

Conclusions

There is clear evidence that not only tobacco smoking but also involuntary exposure to tobacco smoke increases cotinine levels in plasma, with probably a higher increase for former smokers.

Impact

This study supports the evidence that there should be a ban on smoking.

Short running title: Exposure of tobacco smoke and cotinine levels

Keywords: smoking, second hand smoke, cotinine.

List of abbreviations

SHS: of second-hand smoke; BMI: body mass index; NeSNP: never-smoker and no-passive exposure; NeSP: never-smoker and passive exposure; FSNP: former-smoker and no-passive exposure; FSP: former-smoker and passive exposure; CS: current smoker EPIC: European Prospective Investigation into Cancer and Nutrition; GLM: Generalized linear model; BMI: body mass index (in kg/m²).

Introduction

The carcinogenic causing effect of tobacco smoke was established when Doll and Hill conducted the first large epidemiological study in 1950 (1). Today, smoking is recognized as a major contributing factor in a vast variety of conditions including cancers, vascular and respiratory diseases (2, 3, 4), not only to the individual who smokes but to others exposed to second hand tobacco smoke (SHS) (5, 6). Accordingly, in 2002 the IARC Working Group classified SHS as a group I human carcinogen (7, 8). Several countries are discussing new legislation on the ban of smoking in public places: however, the “World Health Organization Report on the Global Tobacco Epidemic, 2009: Implementing smoke-free environments” (9) reports that in 2008 only 17 countries had smoke-free policies that provide universal and effective protection from SHS.

Despite the fact that self-administered questionnaires are frequently limited in terms of reliability and validity, they are routinely used to investigate smoking habits. Gorber et al. (10) have suggested that self-reported estimates may underestimate true smoking prevalence, with higher discrepancies in populations in which smoking is socially undesirable. Therefore, due to the importance of smoking status and exposure to SHS for epidemiological research, biomarkers have been developed. Cotinine is the main metabolite of nicotine and a reliable biomarker of recent nicotine intake (11, 12). Although some foods contain small amounts of nicotine, cotinine level is considered as a reliable indicator of recent exposure to SHS (13).

Cotinine has a half-life of about 15 to 40 hours, reflecting both active and passive smoking during previous days (14). This biomarker has been linked to several modulating factors, including gender, body mass index (BMI), race, number of cigarettes, types of cigarette, filter, and the way the product is smoked (15, 16). Picket et al. (17) have compared serum cotinine levels in non-smokers living under extensive, limited and no coverage smoke-free laws in different locations in the United States. To ensure that the sample included only non-smokers, they used information from self-administered questionnaires and only included subjects who had serum cotinine levels lower or equal to 10 ng/ml (or 56.8 nmol/L). Subjects who declared passive smoke at home were also excluded from the sample. The results show that among non-smoking adults living in counties with extensive smoke-free law coverage, 12.5% were exposed to SHS, compared with 35.1% with limited coverage and 45.9% with no coverage.

In a meta-analysis of studies of passive smoking and lung cancer (18) it was found a pooled relative risk of 1.22 among never-smoking females exposed to spousal smoking, taking into account five cohort studies; the pooled odds ratios were 1.18 and 1.33, respectively, when considering 23 population-based case-control studies and 22 non-population-based case-control studies. The authors highlighted the fact that although the excess risk from SHS exposure is small, the high prevalence of this exposure makes it an important risk factor for lung cancer among non-smoking people.

Vineis et al. (8, 19) previously explored the carcinogenic effects of SHS in the European Prospective Investigation into Cancer and Nutrition (EPIC) study and

observed that non-smokers (never and former smokers) subjected to environmental smoke are at greater risk of lung cancer than subjects with no exposure. They estimated from the EPIC cohort that among never and former smokers the SHS attributable proportion for lung cancer was 16%.

The objective of the present study is to quantitatively assess the relationship between smoking status, SHS and cotinine levels in plasma, using a nested sub cohort within the EPIC cohort.

Methods

The EPIC cohort

The EPIC is an ongoing multi-centre European cohort study which recruited more than 520,000 healthy volunteers from 23 centres in 10 countries (Sweden, Denmark, Norway, the Netherlands, the United Kingdom, France, Germany, Spain, Italy, and Greece) between 1993 and 1998. The study includes men and women with ages mainly between 35 and 74 at recruitment. Dietary and non-dietary (lifestyle, anthropometric) information was collected through two questionnaires at enrolment. Further details have been published previously elsewhere (20).

The current study utilises data from a nested case-control within the EPIC cohort, which includes 899 cases of lung cancer and 1770 controls, matched by country, gender, date of blood collection and date of birth (21). Subjects with missing information for smoking status, cotinine level, sex, age or BMI were excluded from the sample.

Consent was obtained from all participants for the use of their blood samples for future analyses. The study was approved by local ethics committees in the

participating countries and the ethical review board of the International Agency for Research on Cancer (IARC).

Self-reported smoking

Information on SHS at home or at the workplace was collected in 13 out of 23 EPIC centres (6 in France, 5 in Italy, 1 in The Netherlands and 1 in Germany), including home and workplace exposures, using questions such as: “Does someone regularly smoke at home?” and “At work, are there people smoking in your presence?”. The magnitude of SHS exposure was assessed using questions on the number of hours per day of passive smoking each participant was exposed to and the number of cigarettes per day smoked at home and at the workplace.

Laboratory analyses

Participants were also asked to provide a blood sample at recruitment, following a standardised protocol.

Plasma cotinine measurements were available for 2,699 participants within the nested case-control study design. Laboratory analysis was performed at the laboratory of Bevitall AS (Bergen, Norway). Concentration of cotinine was determined by mass spectrometry-based methods (liquid chromatography coupled to tandem mass spectrometry, gas chromatography coupled to tandem mass spectrometry) with sensitivity of 1 nmol/L (0.18 ng/mL) (22).

Statistical analysis

In order to evaluate the effect of smoking status as well as SHS on cotinine levels, subjects were classified according to the following criteria: never-smoker and no-passive exposure (NeSNP), never-smoker and passive exposure (NeSP), former-smoker and no-passive exposure (FSNP), former-smoker and passive exposure (FSP), current smoker (CS) where passive exposure could be at home or at the workplace.

Information on passive exposure (yes or no) was available only for the sample from France, Italy, The Netherlands and Germany. Since the distribution of cotinine was asymmetrical, for our analytical purposes we approximated it to a gamma distribution and transformed the data by adding one unit to all values of plasma cotinine (nmol/L), in order to deal with zero plasma cotinine levels. This only affects the intercept (shifting the curve), not affecting the estimate of the slopes. Then, the association between smoking categories and cotinine levels was assessed using a generalized linear model (GLM) with gamma distribution, adjusting for age, sex, BMI and country (because nicotine contents may differ among brands of cigarettes). BMI was considered as a confounder because subjects with higher BMI tend to exhibit lower cotinine concentrations, possibly because of distribution of cotinine into bigger volume of blood (16). The sample size for this analysis was 859.

We built two further GLMs, one for passive exposure evaluating the effect of hours of SHS per day among non-smokers (never and former). Information about hours of SHS per day was available only in the sample from France and Italy (n=107). A last model was built for smokers evaluating the effect of the average number of cigarettes

currently smoked per day. The sample size was 832 (information available from France, Italy, Spain, United Kingdom, The Netherlands, Greece, Germany and Sweden). In this last model the adjusted model presented a better fit considering the normal distribution.

All statistical analyses were carried out using SAS 9.2 for Windows (SAS Institute, Cary, NC, USA).

Results

Amongst the 3,818 subjects, 2,669 had information on cotinine level and 859 on both cotinine and passive smoke exposure. Some subjects who had declared not being smokers presented very high levels of cotinine. Usually people with cotinine levels higher than 80 nmol/L (14.08 ng/mL) are considered as active smokers (23). We excluded 9 self-declared non-smokers with cotinine levels higher than 500 nmol/L (88.03 ng/mL) (more than 2.65 SD away from the non-smoker mean), considered to be outliers. This exclusion affected participants from the following exposure categories (plasma cotinine levels in nmol/L): 2 NeSNPs (893.23 and 1511.2), 5 FSNPs (1053.1, 2480.85, 1249.13 and 668.63), and 2 FSPs (555.59 and 523.78). We compared five categories of smoking and passive smoking: NeSNP, NeP, FSNP, FSP and CS.

Tobacco exposure categories were found to be significant predictors ($p < 0.0001$) of plasma cotinine levels adjusting for country, sex, age and BMI. Table 1 shows

plasma cotinine means, while Table 2 shows the mean comparisons between pairs of exposure categories.

Amongst non-SHS exposed subjects, there was no significant difference in plasma cotinine concentrations between never and former smokers (mean difference of 0.84 nmol/L, $P=0.29$), whereas the cotinine levels of current smokers were considerably higher than the two other categories of non-active smoking; 1283.85 nmol/L or 226.03 ng/mL higher than never smokers ($P<0.0001$) and 1283.01 nmol/L or 225.88ng/mL higher ($P<0.0001$) than former smokers.

NeSP had cotinine concentrations that were 3.07 nmol/L (0.54 ng/mL) higher than NeSNP. This was similar for former smokers where FSP compared to FSNP presented cotinine concentrations 5.85 nmol/L (1.03 ng/mL) higher ($P = 0.01$). Among those who were passively exposed to cigarette smoke, there was a 3.61 nmol/L difference in cotinine concentrations between never and former smokers that was not statistically significant ($P=0.12$).

In the sample with information on the number of hours of passive smoking ($n=107$, with 2 outliers deleted), an increase of one hour of passive smoking a day was associated with a 2.58 nmol/L (0.45 ng/mL) ($P<0.001$) higher mean plasma cotinine concentration after controlling for country, age, sex and BMI (Table 3). Among current smokers ($n=832$), an increase of one cigarette per day was associated with a 22.44 nmol/L (3.95 ng/mL) ($P<0.001$) higher average plasma cotinine level after controlling for country, age, sex and BMI (Table 4).

As reported in Table 1, cotinine levels are much higher for smokers. In Table 4 it is possible to estimate the expected level of cotinine for one person with some characteristics regarding the covariates. For example, one subject from Greece

female, 40 years, BMI of 22, who smokes one cigarette per day, presents an expected value of 1427.40 nmol/L (251.3 ng/mL) of cotinine levels. If she smokes two cigarettes per day, her expected level will increase to 1449.83 nmol/L (25.25 ng/mL). Or if it was a men, these values would be 1536.49 nmol/L (270.5 ng/mL) and 1558.92 (274.5 ng/mL), respectively.

Discussion

In this study, we have applied three linear models to investigate how components of smoking exposure are quantitatively related to plasma cotinine levels (15, 16, 24). The first model showed that among non-smokers without SHS exposure, cotinine levels did not differ between never and former smokers. Among those who were passively exposed, FSP had slightly higher cotinine concentrations on average when compared to NeSP, but the difference was not significant ($P=0.12$). The sample size of the FSP group is quite small ($n=37$) with average cotinine levels of 9.66 nmol/L and a large standard error (2.16 nmol/L). The 95% confidence interval presents a lower limit of 5.95 nmol/L which is quite lower than the upper limit 8.15 nmol/L of the NeSP group. The number of passive smoking hours are similar between the two categories NeSP and FSP (means of 1.28 and 1.17 hours respectively), but it should be noted that the sample size for former smokers with passive smoking hours information is limited ($n=6$ data not shown). In a recent paper Piccardo et al (25) suggest that among smokers the exposure to their own environmental smoke is not negligible, if smoking occurs in indoor environments. We did not compare cotinine

levels among current smokers also exposed to passive smoking (with and without passive exposure).

In our results it is clear that amongst non-smokers, passive exposure significantly increases cotinine levels (never smoker, $P < 0.0001$ and former, $P = 0.01$). In general, one extra hour of passive exposure means a rise of 2.58 nmol/L (0.45 ng/mL) in cotinine levels. A range of studies show that cotinine levels increase with an increasing number of hours of tobacco exposure (26) or with the number of cigarettes smoked (10, 15, 16, 23), but there is no study comparing never and former smokers. Whincup et al (6) found an excess of risk of coronary heart disease for higher concentrations of serum cotinine among non-smokers. Gallo et al (5) showed that SHS at home increases the risk of dying from cardiovascular diseases.

The increase in cotinine levels between those non-exposed to SHS and those exposed was 3.07 nmol/L (0.54 ng/mL) for never smokers and 5.85 nmol/L (1.03 ng/mL) for former smokers. Vineis et al (8) suggested that former smokers may be more susceptible to the effects of environmental tobacco smoke, having higher levels of cotinine when compared to never smokers, after controlling for number of hours of passive smoking. The fact that cotinine levels were very similar between former and never smokers suggests that cotinine levels reduce upon cessation of smoking due to the short half life. However, we do not have data for cotinine levels by time since cessation, which is a limitation of the study.

Our findings on current smokers concur with previous publications: cotinine levels increase when the mean number of cigarettes per day increases (15, 16, 24). As

shown in Table 4, the magnitude of rise was 22.44 nmol/L (3.95 ng/mL) for each increase in one cigarette smoked per day.

Our results regarding the relations between cotinine levels and categories of smoking status, cigarettes per day and exposure to SHS show that the questions related to reported tobacco exposure are accurate.

Strulovici-Barel et al (27) evaluated bronchoscopy and genome wide gene expression of 121 healthy volunteers from New York City. The group was separated into non-smokers, active smokers and individuals exposed to low levels of tobacco smoking (environmental tobacco exposure and/or occasional smoking). The results showed that the small airway epithelium is very sensitive to low level tobacco smoke exposure (nicotine and cotinine levels were considered for this classification). They conclude that “the changes in gene expression are likely the earliest biological abnormalities in the small airway epithelium that lead to clinically detectable lung disease in some individuals”. This finding gives support to our hypothesis that the higher levels of cotinine concentrations, even if not as high as smokers’ levels, may increase the risk at least of lung diseases.

Conclusions

We conclude that there is evidence that not only active tobacco smoking but involuntary exposure increases plasma cotinine levels. Never smokers are exposed to higher levels of cotinine when exposed to passive smoke and former smokers may be exposed to an even higher level. We strongly support the smoking ban laws which ensure protection to never smokers and former smokers.

References

1. Doll R, Hill A. Smoking and carcinoma of the lung: preliminary report. *Br Med J* 1950; **2**(4682): 739-48.
2. Gan Q, Smith KR, Hammond SK, Hu T-w. Disease burden of adult lung cancer and ischemic heart disease from passive tobacco smoking in China. *Tobacco control* 2007; **16**:417-422.
3. López MJ, Pérez-Ríos M, Schiaffino A, et al. Mortality attributable to passive smoking in Spain, 2002. *Tobacco control* 2007; **16**:373-377.
4. Vineis P, Hoek G, Krzyzanowski M, et al. Lung cancers attributable to environmental tobacco smoke and air pollution in non-smokers in different European countries: a prospective study. *Environmental Health* 2007; **6**:7.
5. Gallo V, Neasham D, Airoidi L, et al. Second-hand smoke, cotinine levels, and risk of circulatory mortality in a large cohort study of never-smokers. *Epidemiology*. 2010; **21**(2):207-14.
6. Whincup PH, Gilg JA, Emberson JR, et al. Passive smoking and risk of coronary heart disease and stroke: prospective study with cotinine measurement. *BMJ*. 2004; **329**(7459):200–205.
7. Tobacco Smoking and involuntary smoking. *IARC Monographs on the evaluation of Carcinogenic risks to Humans*. Vol. 83. Lyon; 2004.
8. Vineis P, Airoidi L, Veglia P, et al. Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ* 2005; **330**(7486):277-81.
9. WHO report on the global tobacco epidemic, 2009: implementing smoke-free environments. *World Health Organization, WHO Press*, Geneva: 2009.

10. Gorber SC, Schofield-Hurwitz S, Hardt J, et al. The accuracy of self-reported smoking: A systematic review of the relationship between self-reported and cotinine assessed smoking status. *Nicotine Tob Res.* 2009; **11(1)**:12-24.
11. Perezstable EJ, Benowitz NL, Marin G. Is Serum Cotinine a Better Measure of Cigarette-Smoking Than Self-Report? *Prev Med.* 1995; **24(2)**:171-9.
12. Benowitz NL. Cotinine as a Biomarker of Environmental Tobacco Smoke Exposure. *Epidemiologic Reviews* 1996; **18 (2)**: 188-204
13. Benowitz NL. Biomarkers of environmental tobacco smoke exposure. *Environ Health Perspect* 1999; **107 (2)**:349-55.
14. Benowitz NL, Schultz KE, Haller CA, et al. Prevalence of Smoking Assessed Biochemically in an Urban Public Hospital: A Rationale for Routine Cotinine Screening. *Am J Epidemiol.* 2009; **170(7)**: 885-91.
15. Figueiredo VC, Szklo M, Szklo AS, et al. Determinants of salivary cotinine level: a population-based study in Brazil. *Rev Saude Publica.* 2007;**41(6)**: 954-62.
16. Jain RB, Bernert JT. Effect of body mass index and total blood volume on serum cotinine levels among cigarette smokers: NHANES 1999-2008. *Clin Chim Acta.* 2010; **411(15-16)**:1063-8.
17. Pickett MS, Schober SE, Brody DJ, et al. Smoke-free laws and secondhand smoke exposure in US non-smoking adults, 1999-2002. *Tobacco Control* 2006; **15**: 302-307.
18. Taylor R, Najafi F, Dobson A. Meta-analysis of studies of passive smoking and lung cancer: effects of study type and continent. *Int J of Epidemiol* 2007; **36**:1048–1059.
19. Vineis P, Alavanja M, Buffler P, et al. Tobacco and cancer: recent epidemiological evidence. *J Natl Cancer Inst* 2004; **96(2)**:99-106.
20. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002, **5(6B)**:1113-1124
21. Johansson M, Relton C, Ueland PM, et al. Serum B vitamin levels and risk of lung cancer. *JAMA.* 2010; 303(23):2377-2385.

22. Ueland PM, Midttun O, Windelberg A, et al. Quantitative profiling of folate and one-carbon metabolism in large-scale epidemiological studies by mass spectrometry. *Clin Chem Lab Med*. 2007; **45(12)**: 1737-1745.
23. Jarvis MJ, Tunstall-Pedoe H, Feyerabend C, et al. Comparison of tests used to distinguish smokers from nonsmokers. *Am J Public Health*. 1987;**77**:1435–8
24. Gan WQ, Cohen SB, Man SF, et al. Sex-related differences in serum cotinine concentrations in daily cigarette smokers. *Nicotine Tob Res*. 2008; **10(8)**:1293-300.
25. Piccardo MT, Stella A, Valerio F. Is the smokers exposure to environmental tobacco smoke negligible? *Environmental Health* 2010; 9:5 (doi:10.1186/1476-069X-9-5).
26. Jarvis MJ, Feyerabend C, Bryant A, et al. Passive smoking in the home: plasma cotinine concentrations in non-smokers with smoking partners. *Tobacco Control* 2001; **10**:368–374.
27. Strulivici-Barel Y, Omberg L, O'Mahony M, et al. Threshold of biologic responses of the small airway Epithelium to low levels of tobacco smoke. *AJRCCM Article in press* (August 2010).

Table 1. Descriptive statistics for cotinine* by SHS categories (n=859), controlling for country, age, sex and BMI.

Category	Adjusted Cotinine				
	N n =859	Mean (nmol/L)	Standard Error	IC 95%	
NeSNP	125	2.98	0.39	3.21	4.74
NeSP	129	6.05	0.56	5.95	8.15
FSNP	114	3.81	0.67	3.51	6.12
FSP	37	9.66	2.16	6.42	14.90
CS	454	1286.83	53.76	1182.45	1393.20

*to convert nmol/L to ng/mL, divide cotinine mean by the factor 5.68

Table 2. Mean comparison between pairs of smoking categories (model 1).

Adjusted Mean difference* (column-row) (P-value) (CI95 %)				
	NeSP	FSNP	FSP	CS
	3.07 (P<0.001)	0.84 (P=0.29)	6.68 (P=0.003)	1283.85 (P<0.001)
NeSNP	(1.92; 4.22)	(-0.71; 2.39)	(2.20; 11.16)	(1178.47; 1389.23)
		-2.24 (P=0.02)	3.61 (P=0.12)	1280.7 (P<0.001)
NeSP		(-4.06; -0.41)	(-0.94; 8.15)	(1175.39; 1386.16)
			5.85 (P=0.01)	1283.01 (P<0.001)
FSNP			(1.37; 10.32)	(1177.63; 1388.39)
				1277.16 (P<0.001)
FSP				(1171.70; 1382.63)

* GLM model controlling for country (P=0.34) sex (b=1.44, P=0.04), age (b=-0.025, P=0.41) and BMI (b=0.19, P=0.008), with Gamma distribution

*to convert nmol/L to ng/mL, divide cotinine mean by the factor 5.68

Table 3. Generalized Linear Models for non-smokers (never and former) exposed to SHS.

Model 2**(n=107)		Estimate*	Standard Error	Wald 95% Confidence Limits		p-value
No-smoker	Hours/day of passive smoking	2.58	0.60	1.41	3.75	<0.001
Country (France/Italy)	(Italy as baseline)	0.18	1.00	-1.79	2.15	0.859
Sex	Male/Female	0.51	0.83	-1.12	2.14	0.514
age	years	-0.01	0.05	-0.12	0.10	0.883
BMI	Kg/m2	-0.15	0.14	-0.42	0.13	0.299

* per unit increase

** GLM with gamma distribution

*** Intercept of 7.04 (-1.64; 15.73)

****to convert nmol/L to ng/mL, divide cotinine mean by the factor 5.68

Table 4. Generalized Linear Models for smokers.

Model 3***(n=832)		Estimate*	Standard Error	Wald 95% Confidence Limits		p-value
Smokers	per cigarettes/day	22.44	1.81	18.90	25.98	<0.001
Country (Greece as baseline)	France/Greece	26.81	163.47	-293.58	347.20	0.870
	Italy/Greece	149.85	76.67	-0.43	300.12	0.051
	Spain/Greece	80.39	75.42	-67.43	228.20	0.287
	UK/Greece	158.98	85.99	-9.56	327.53	0.065
	NL/Greece	269.80	79.96	113.08	426.52	0.001
	Germany/Greece	206.63	71.87	65.76	347.49	0.004
Sex	Male/Female	109.09	50.42	10.26	207.92	0.031
age	years	-2.22	2.70	-7.51	3.06	0.410
BMI	Kg/m2	-31.88	4.89	-41.48	-22.29	<0.001

* per unit increase

** GLM with normal distribution

***Intercept of 2195.276 (1794.87; 2595.68)

4.2.MANUSCRITO 2

Artigo submetido no periódico *American Journal of Epidemiology*.

The role of one-carbon metabolism in lung cancer aetiology: disentangling metabolic pathways with structural equation models

Valéria Troncoso Baltar^{1§}, Mattias Johansson², Wei Xun³, Pietro Ferrari², Shu-Chun Chuang³, Caroline Relton⁴, Per Magne Ueland^{5,6}, Øivind Midttun⁷, Nadia Slimani², Mazda Jenab², Françoise Clavel-Chapelon^{8,9,10}, Marie-Christine Boutron-Ruault^{8,9,10}, Guy Fagherazzi^{8,9,10}, Rudolf Kaaks¹¹, Sabine Rohrmann¹², Heiner Boeing¹³, Cornelia Weikert¹³, H. Bas Bueno-de-Mesquita¹⁴, Hendriek Boshuizen¹⁴, Carla H. van Gils¹⁵, N Charlotte Onland-Moret¹⁵, Antonio Agudo¹⁶, Aurelio Barricarte^{17,21}, Carmen Navarro^{18,21}, Laudina Rodríguez¹⁹, José Maria Huerta Castaño^{20,21}, Nerea Larrañaga^{21,22}, Kay-Tee Khaw²³, Nick Wareham²⁴, Naomi E. Allen²⁵, Francesca Crowe²⁵, Valentina Gallo³, Teresa Norat³, Vittorio Krogh²⁶, Giovanna Masala²⁷, Salvatore Panico²⁸, Carlota Sacerdote²⁹, Rosario Tumino³⁰, Antonia Trichopoulou^{31,32}, Pagona Lagiou^{31,33}, Dimitrios Trichopoulos^{33,34}, Torgny Rasmuson³⁵, Göran Hallmans³⁶, Nina Roswall³⁷, Anne Tjønneland³⁷, Elio Riboli³, Paul Brennan², Paolo Vineis^{30,38}

¹Department of Epidemiology, Faculty of Public Health, University of São Paulo, Av. Dr Arnaldo, 715. CEP 01246-904. São Paulo, São Paulo, Brazil

²International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372, Lyon, France

³Department of Epidemiology & Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, St Mary's Campus, 5th floor, Norfolk Place W2 1PG, London, UK

⁴Institute for Ageing and Health, Newcastle University, Campus for Ageing and Vitality, Newcastle upon Tyne, NE4 5PL Newcastle, UK

⁵Section for Pharmacology, Institute of Medicine, University of Bergen, 9th floor, New Laboratory Building 5021, Bergen, Norway

⁶Haukeland University Hospital, Jonas Lies vei 65 N-5058. PO.Box 1, N-5021, Bergen, Norway

⁷Bevital AS, Laboratoriebygget, 9 etg., 5021 Bergen, Norway

⁸Institut National de la Santé et de la Recherche Médicale, INSERM U780-IFR169, 16 Avenue Paul Vaillant-Couturier, 94807 Villejuif, France

⁹Institut Gustave Roussy, 39 rue Camille Desmoulins, 94805, Villejuif, France

¹⁰Paris South University, Villejuif, France

¹¹Division of Cancer Epidemiology, German Cancer Research Centre, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

¹²Division of Clinical Epidemiology, German Cancer Research Centre, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

¹³Department of Epidemiology, German Institute of Human Nutrition, Arthur-Scheunert-Allee 114-116, 14558 Bergholz-Rehbrücke, Nuthetal, Germany

¹⁴National Institute of Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, The Netherlands

¹⁵Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Room Str. 6.131, PO Box 85500, 3508 GA Utrecht, The Netherlands

¹⁶Unit of Nutrition, Environment and Cancer, Catalan Institute of Oncology, Gran Via 199-203, 08907 L'Hospitalet de Llobregat, Spain

¹⁷Epidemiology, Prevention and Promotion Health Service, Public Health Institute of Navarra, Leyre 15, 31003 Pamplona, Navarra, Spain

¹⁸Department of Epidemiology, Regional Council of Health and Consumer Affairs, Ronda de Levante 11, Murcia 3008, Spain

¹⁹Public Health and Participation Directorate, Health and Health Care Services Council, C/ Ciriaco Miguel Virgil no 9, CP 33006 Oviedo, Asturias, Spain

²⁰Department of Epidemiology, Murcia Regional Health Authority, Ronda de Levante 11, Murcia 3008, Spain

²¹CIBER of Epidemiology and Public Health, C/ Dr. Aiguader, 88, 1a planta – 08003, Barcelona, Spain

²²Epidemiology Unit, Public Health Department of Gipuzkoa, Basque Government Avenida de Navarra, 4-20013 Donostia, San Sebastián, Spain

²³Cancer registry of anada, Andalusian School of Public Health, Cuesta del observatorio, 4, 18080, Granada, Spain

²⁴Clinical Gerontology Unit, Department of Public Health and Primary Care, University of Cambridge, Level 2, F + G Block, Box 251, Addenbrooke's Hospital Hills Road, CB2 2QQ Cambridge, UK

²⁵MRC Epidemiology Unit, Strangeways Research Laboratory, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK

²⁶Cancer Epidemiology Unit, University of Oxford, Richard Doll Building, Old Road Campus, OX3 7LF Oxford, UK

²⁷Environmental Epidemiology and Cancer Registry Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, via Venezian , 1, I - 20133 Milano , Italia

²⁸Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Centre, Scientific Institute of Tuscany, Via Cosimo il Vecchio 2, 50139 Florence, Italy

²⁹Department of Clinical and Experimental Medicine, Federico II University of Naples, Via Pansini 5, 80131 Naples, Italy

³⁰Human Genetic Foundation (HuGeF), 52, via Nizza 10123 Torino Italy,

³¹Ragusa Cancer Registry, Azienda Ospedaliera "Civile M.P. Arezzo", Via Dante N° 109, 97100 Ragusa, Italy

³²WHO Collaborating Center for Food and Nutrition Policies, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School 75 M. Asias Avenue, Goudi, GR-115 27, Athens, Greece

³³Hellenic Health Foundation, 10-12 Tetrapoleos Street, GR-115 27, Athens Greece

³⁴Department of Epidemiology, Harvard School of Public Health, Harvard University
677 Huntington Ave, Boston, MA 02115. USA

³⁵Department of Radiation Sciences, Oncology, Umeå University, SE-901 87 Umeå,
Sweden

³⁶Department of Public Health and Clinical Medicine, Umeå University, SE-90187 ,
Sweden

³⁷Institute of Cancer Epidemiology, Danish Cancer Society, Strandboulevarden 49,
2100, Copenhagen, Denmark

³⁸Department of Epidemiology & Public Health, MRC-HPA Centre for Environment
and Health, Imperial College, Norfolk Place, W2 1PG, London, UK

[§]Corresponding author

Email addresses: VTB: vbaltar@usp.br, MJ: johanssonm@fellows.iarc.fr, WX: wei.xun06@imperial.ac.uk, PF: pietro2204@yahoo.com, SCC: s-c.chuang@imperial.ac.uk, CR: c.l.relton@newcastle.ac.uk, PMU: per.ueland@ikb.uib.no, ØM: bjorn.midttun@farm.uib.no, NS: slimani@iarc.fr, MJ: jenab@iarc.fr, FC-C: clavel@igr.fr, M-CB-R: boutron@igr.fr, GF: guy.fagherazzi@igr.fr, RK: r.kaaks@dkfz-heidelberg.de, SR: sabine.rohrmann@ifspm.uzh.ch, HB: boeing@mail.dife.de, CW: weikert@dife.de, HBBd-M: Bas.Bueno.De.Mesquita@rivm.nl, HCB: Hendriek.Boshuizen@rivm.nl, CHvG: C.vanGils@umcutrecht.nl, PP: P.H.M.Peeters@umcutrecht.nl, AA: a.agudo@iconcologia.net, AB: abarricg@cfnavarra.es, CN: carmen.navarro@carm.es, LR: laudina.rodriquezsuares@asturias.org, JMHC: jmhuerta.carm@gmail.com, NL: epidem3-san@ej-gv.es, MJSP: mariajose.sanchez.easp@juntadeandalucia.es, K-TK: kk101@medschl.cam.ac.uk, NW: nick.wareham@mrc-epid.cam.ac.uk, NA: Naomi.allen@ceu.ox.ac.uk, FC: francesca.crowe@ceu.ox.ac.uk, VG: v.gallo@imperial.ac.uk, TN: t.norat@imperial.ac.uk, GT: giovanna.tagliabue@istitutotumori.mi.it, GM: g.masala@ispo.toscana.it, SP: spanico@unina.it, CS: carlotta.sacerdote@cpo.it, RT: rtumino@tin.it, AT: atrichopoulou@hhf-greece.gr, PL: pdlagiou@med.uoa.gr, CB: cbamia@nut.uoa.gr, TR: torgny.rasmuson@onkologi.umu.se, GH: goran.hallmans@nutrires.umu.se, NR: roswall@cancer.dk, AT: annet@cancer.dk, ER: e.riboli@imperial.ac.uk, PB: brennan@iarc.fr, PV: p.vineis@imperial.ac.uk

ABBREVIATIONS:

AIC: Akaike information criterion; β , linear regression coefficient; CI, confidence interval; EPIC, European Prospective Investigation into Cancer and Nutrition; KRT, kynurenin/tryptophan ratio; log2, base 2 logarithm; MTHFR, methylenetetrahydrofolate reductase; OR, odds ratio; OCM, one-carbon metabolism; *P*, *P* value; SAM, S-adenosylmethionine; SD, standard deviation; SEM, structural equation model; TMP, thymidylate; 5-methylTHF, 5-methyltetrahydrofolate; 5,10-methyleneTHF, 5,10-methylenetetrahydrofolate;

SUMMARY

Background: The one-carbon metabolism (OCM) is considered key in maintaining DNA integrity and regulating gene expression, and may thus affect carcinogenesis. Several components of OCM were recently implicated in lung cancer risk in a nested case-control study within the European Prospective Investigation into Cancer and Nutrition (EPIC). It is unclear whether these factors act independently or through common mechanisms. **Methods:** The current study applies structural equations modelling (SEM) to further disentangle the mechanisms involved in lung carcinogenesis. SEM allows simultaneous estimation of linear relations where variables can be dependent in one equation and independent in another. Biomarkers were analysed from 891 lung cancer cases and 1,747 controls nested within EPIC. Three putative mechanisms in the OCM were investigated in relation to lung cancer risk: methylation, DNA synthesis and transsulfuration, as well as mechanisms involved in immune activation, always controlling for cotinine levels (recent tobacco exposure). Analyses were performed for base 2 logarithm (log) and standardized scales (SD). **Results:** The best model fit was observed for the model which includes only methylation, modelled through methionine, leading to an odds ratio of 0.52 (95% confidence interval = 0.39; 0.69) per doubling in serum levels. Furthermore, cobalamin, vitamins B6 and folate seem to participate in this mechanism with indirect effects. **Conclusion:** The results of this study are consistent with the evidence that one-carbon metabolism affects lung cancer risk mainly through methylation processes.

Keywords: B vitamins, folate, methionine, lung cancer, inflammation, structural equation model.

INTRODUCTION

Lung cancer continues to be a major public health problem world-wide. The World Health Organization estimated that in 2004 1.3 million deaths were due to cancer of the lung, out of 7.4 million deaths from all cancers (1). In addition to exposures to carcinogens or carcinogenic mixtures such as smoking and air pollution, low-micronutrient intake has been linked to lung cancer (2, 3).

Amongst the micronutrients that have been investigated, some B vitamins and other components within the pathways of one-carbon metabolism (OCM) have been associated with carcinogenesis (3). The OCM describes a network of several interlinked reactions in which folate (vitamin B9) acts as a crucial carrier of methyl (one-carbon) groups. The pathways of OCM include several water-soluble B vitamins such as folate, vitamins B6 and B12, as well as amino acids such as homocysteine, methionine and serine (Figure 1), most of which are derived from diet. One branch of the metabolism is the synthesis of methionine and S-adenosylmethionine (SAM) which is essential for DNA methylation. This is modulated by the intake of dietary folate, which acts as a carrier of single-carbon units, essential for both the biosynthesis of nucleotides such as purines and pyrimidine, as well as in the process of remethylation of homocysteine to methionine (4-8). Folate is metabolised to 5,10-methylenetetrahydrofolate (5,10-methyleneTHF), which is subsequently converted by the enzyme methylenetetrahydrofolate reductase (MTHFR) to 5-methyltetrahydrofolate (5-methylTHF), by adding a methyl group to homocysteine to produce methionine and SAM. Vitamin B12 is essential in this conversion (9). Methionine is an amino acid which is converted to S-adenosylmethionine (SAM), an essential compound for

DNA methylation (4). Folate or vitamin B12 deficiency results in decreased SAM which, in turn, leads to elevated homocysteine levels (6), due to decreased methionine synthesis. Finally, methionine and SAM, together with the enzyme DNA methyltransferase, lead to DNA methylation. During DNA methylation some methionine is re-used for homocysteine remethylation (10).

There are two primary mechanisms by which OCM has been proposed to influence cancer development: (i) DNA methylation, and (ii) DNA synthesis and repair (4, 6). Gene-specific hypermethylation and global hypomethylation are frequently observed in many types of tumour tissue, while folate deficiency can cause thymidylate (TMP) stress during DNA synthesis, thus resulting in DNA strand breaks and genomic instability (4, 9). The addition of a methyl group at the carbon-5 position of the cytosine ring is termed DNA methylation, a very common epigenetic event. DNA methylation is an important regulatory mechanism in gene transcription and can initiate a cascade of processes which have been implicated in a variety of tumours (9), heart diseases and neural tube defects (9). Several genes are reportedly hyper-methylated in human lung cancer including APC, CDKN2A/p16, DAPK1, MGMT and RASSF1A (9, 11, 12).

Vitamin B6 functions act as a co-factor in more than 100 reactions in the body (13), including both the transsulfuration pathway (14), as well as several steps of the kynurenine pathway of tryptophan metabolism that are related to inflammation and immune activation (15).

In a recent case-control study of lung cancer nested within the EPIC study, strong protective effects were observed for increasing serum levels of both vitamin B6 and methionine, independently of tobacco smoking and other potential

confounding factors (6). In mutually adjusted logistic regression analysis these effects appeared as independent, but it is possible that they interact in a more complicated way in their relation with risk, which is not straightforwardly captured by standard regression analysis.

Using structural equation models (SEM) it is possible to evaluate pathways involving multiple intermediate factors. In order to further investigate the complex network of pathways through which OCM components might influence lung cancer risk, we have conducted a study using the same dataset as in the original EPIC study (6). We have hypothesized a model of the OCM pathway including serum measurements of several key factors and evaluated their association with risk. Care has been taken to control for potential confounders, particularly tobacco smoking, and serum cotinine as a biomarker of short-term nicotine exposure (16) was included in the model.

MATERIALS AND METHODS

Study subjects

The European Prospective Investigation into Cancer and Nutrition (EPIC) study is a multi-centre study involving 10 European Countries and 23 centres (17). Among the 519,978 subjects, 2,206 were diagnosed with lung cancer after an average of 10 years of follow-up. The lung cancer cases were defined on the basis of the International Classification of Diseases for Oncology, Second Edition, and included all invasive cancers coded as C34. Cases with no available blood sample or date of blood collection or history of another cancer at the time of blood donation were excluded. For each case, two controls were randomly matched considering: country, gender,

date of blood collection and date of birth (6). Cases with no controls were excluded from the sample. From the nested case-control, 39 subjects were excluded because of missing information from laboratory analyses or because their matched counterpart(s) had missing information. The final data, with complete information on folate, serine, vitamins B6 and B12, cotinine, homocysteine, methionine, cysteine, cystathionine, neopterin, and kynurenin/tryptophan ratio (KTR), has 2638 subjects (891 cases and 1747 controls: 856 with 1:2 matching and 35 cases with 1:1).

Biochemical analyses

Measurements of serum concentrations (plasma in Swedish samples) of serine, glycine, vitamin B6, folate, vitamin B12, total homocysteine, methionine, cysteine, cystathionine, neopterin, kynurenin, tryptophan and plasma cotinine were performed at Bevit A/S (www.bevital.no), blinded as to the case-control status (6).

Model assignment and statistical methods

In biological pathways, as in the case of the OCM, determinants of intermediate factors can be considered outcomes in their own right. This is often difficult to handle using traditional regression analysis. SEM can be applied to address such complexities as it allows simultaneous estimation of the linear relations among a number of covariates. In this context, the same variable can be dependent in one equation and independent in another (18). SEMs are frequently used to confirm an *a priori* hypothesis based on a theoretical model, although they can also be used for exploratory analysis. Pathway diagrams are used to help researchers to visualize the theoretical model, each arrow representing one causal relationship.

Initially we built a simplified model A of the OCM pathway (Figure 2, in red), assuming its effect on lung cancer risk is mediated through methylation processes only. In this model, we hypothesised that methionine was the main modifier of lung cancer risk, while homocysteine and methionine were a cause and effect of each other through the re-methylation pathway. The reduced form of folate, 5-tetrahydrofolate, as well as vitamin B12 were linked to methionine levels as they are both required for methionine production involving the enzyme methionine synthase (19, 20). In addition, serine was linked to homocysteine due to its role in transferring the methyl group from SAM to form sarcosine and homocysteine that will be re-methylated (10).

Next, we built model B, adding the nucleotide synthesis branch of the OCM, by allowing folate to directly affect lung cancer status (Figure 2, green plus red). The folate cycle is essential for DNA biosynthesis and cell replication (19), and its cofactors serine and vitamins B2 were included in this branch of the pathway.

We also built model C, assuming that the transsulfuration pathway in which homocysteine is metabolized to cysteine through the sequential reaction using two vitamin B6 dependent enzymes (20), can also directly influence lung cancer risk (Figure 2, blue plus red plus green). This is achieved by adding the transsulfuration pathway onto model B, represented by the direct link from homocysteine to cystathionine, which was then linked to cysteine, with serine and vitamin B6 as cofactors. Finally, cysteine was allowed to directly affect lung cancer risk.

In SEM model D, immune activation and inflammatory activities, modelled through vitamin B6, neopterin and kynurenin/tryptophan ratio (KTR), which may affect lung cancer risk (Figure 2, purple plus red plus green plus blue).

Finally we built model E, in which immune activation and inflammation were modelled through vitamin B6 (with indirect effect of KTR and neopterin), while the transsulfuration branch was excluded. Methylation and nucleotide synthesis were maintained in this model.

Based on a theoretical pathway (Figure 2), all models were identifiable which means that they provided a unique set of parameter estimates. As SEM estimation is based on correlation matrix, *Pearson* correlation was considered between continuous variables and *biserial* correlation was considered between continuous and binary variables. For the final outcome, lung cancer status, the matching design was included by using a complex sample with stratum, which allows the estimation of conditional logistic regression for this outcome (21, 22). Furthermore, all regressions were estimated by maximum likelihood with robust standard errors and chi-square (due to the fact there were continuous and binary outcomes). It was estimated for all models the model selecting Akaike information criterion (AIC) (23).

Due to the strong association between tobacco smoking and lung cancer as well as the effects of tobacco on serum levels of OCM intermediates, including folate and homocysteine, and to a lesser extent methionine (6, 10), serum cotinine (measured in plasma at the same time as serum B vitamins) was included in all equations. All linear regression models (equations 2 to 7) were also adjusted for sex and age; continuous variables were estimated both in their standardized forms (i.e. subtracting the sample mean for each variable and then dividing by the standard deviation), as well as for base 2 logarithms (with the exceptions of cotinine, which was maintained in its standardized form due to zero values, and sex, which did not change). We provide estimates using both transformations. Extensive analyses on the

main effects of the biomarkers on lung cancer status have been published previously by Johansson et al (6).

In order to evaluate model goodness of fit, residual analysis was used plotting Cook's distance and log-likelihood distance influence measures (graphs not shown). It was considered that all models presented good fit because the model coefficients, with and without influent points, were similar. All SEM modelling was performed using Mplus version 6 (22).

RESULTS

Cohort characteristics

In the current nested case-control study, the mean age was 58 years for both cases and controls. Amongst controls, 38.7% were never smokers and 37.0% were former smokers at the time of recruitment, compared to 10.8% and 28.8% respectively among the cases. The mean cotinine level was 231.07 nmol/L for controls and 896.43 nmol/L for cases. Summary information (mean and standard deviation (SD)) for homocysteine, methionine, folate, vitamins B12, B6, B2, cysteine, cystathionine, neopterin and KTR can be found in Table 1.

Disentangling the pathways

Methylation pathway only. To test the *a priori* assumption that the OCM's influence on lung cancer risk is mediated solely by the methylation branch according to model A (Figure 2, in red), three equations were used to describe this pathway, with methionine being the final intermediate factor affecting risk, further modulated by homocysteine, folate, vitamin B12, vitamin B6 and indirectly by serine (Figure 2,

Table 2). Folate was considered as an independent variable in relation to methionine, and was kept in the models due to *a priori* knowledge despite not showing a significant effect in relation to methionine in the analysis of non-log transformed variables. It was also important to include methionine and homocysteine as a loop, since they can exert mutual influence on each other, representing methylation and demethylation reactions. An increase in one SD of methionine or a doubling in dose of methionine was associated with a decreased lung cancer risk of 19% ($P < 0.01$) and 48% ($P < 0.01$), respectively. Homocysteine, vitamin B12 and folate were positively related to methionine, and the re-methylation loop between methionine and homocysteine showed significant associations in both directions, i.e. an increase in homocysteine gave an increase in methionine, whereas an increase in methionine led to a decrease in homocysteine. Interestingly, folate showed an indirect effect on the log2 scale (a doubling in folate was associated with 20% reduction in lung cancer risk via methionine) and vitamin B6 showed an indirect effect on the SD scale (one SD increase was associated with 1% reduction in lung cancer risk). All results for Model A are presented in Table 2.

Methylation and nucleotide synthesis pathways. In order to investigate nucleotide synthesis in relation to risk in addition to the methylation branch of the OCM, we built model B (Figure 2, red plus green colour). This model allows folate to be directly linked to lung cancer status, with folate being modelled by vitamin B2 and serine. Model B (equation 1) suggests that methionine and folate are both directly associated with reduced lung cancer risk through their respective pathways. The decrease in lung cancer risk for an increase of one SD or a doubling in methionine levels was 19% ($P < 0.01$) and 48% ($P < 0.01$), and an increase of one SD or a doubling

in folate levels was associated with a 13% and 12% reduction ($P=0.02$ and $P=0.03$ respectively). No change was seen in the direct or indirect structural relationships for the methionine-homocysteine re-methylation pathway and its components when including the nucleotide synthesis in the SEM. In the DNA synthesis pathway, represented by equation 4, serine was inversely associated with folate, a doubling resulting in a 12% reduction in folate ($1-2^{-0.19}=0.12$), whereas this relation was not significant in the analysis of non-log transformed variables. An increase of one SD or a doubling in vitamin B2 was associated with 0.12 SD and a 14% ($2^{0.20}-1=0.14$) increase in folate, respectively. It is interesting to notice that folate presented a direct and indirect effect on lung cancer risk in the log2 model. The indirect effect via methionine indicates that a doubling in folate was associated with an indirect reduction of 4% in lung cancer risk (a total effect of folate is about a $12+4=16\%$ reduction). All results for model B are presented in Table 2.

Methylation, nucleotide synthesis and transsulfuration pathways. SEM model C (Figure 2, plus blue colour) included the transsulfuration branch, in addition to the methylation (red) and nucleotide synthesis (green) branches. In Figure 2 there are two additional equations representing the transsulfuration branch (equations 5 and 6). Despite the fact that all components of the transsulfuration branch were significantly related with cystathionine and cysteine, they did not affect lung cancer risk because the final mediation endpoint of the transsulfuration branch, here represented by cysteine, was not associated with lung cancer risk ($P=0.30$ for both scales).

Methylation, nucleotide synthesis, transsulfuration and anti-inflammatory pathways. SEM model D adds indicators of the inflammation and immune activation branch to model C. Neopterin is inversely related to vitamin B6 on both scales, whereas KTR

is inversely related only on the log2 scale. In this analysis, on log2 scale, apart from methionine, only vitamin B6 was directly related to lung cancer, a doubling being associated with a 15% reduction in risk ($P=0.01$). This vitamin B6 effect was not detected in the analysis of non-log transformed variables, but vitamin B6 presented an indirect effect via methionine which was not significant on the log2 scale analysis. In addition, on the SD scale, folate presented only a direct effect on lung cancer risk while vitamin B6 had only an indirect effect via methionine. On the log2 scale, folate did not show a direct effect, but showed an indirect effect, while vitamin B6 presented only a direct effect.

Methylation, nucleotide synthesis, and anti-inflammatory pathways. Finally, SEM model E included only three direct pathways to lung cancer, in which the anti-inflammatory branch is represented by vitamin B6 as a final intermediate marker before lung cancer risk (KTR and neopterin direct effects were excluded because they were not significant). Considering the SD scale, folate showed only a direct and protective effect on lung cancer risk, whereas vitamin B6 showed only an indirect effect through the methylation pathway. On the other hand, on the log2 scale, vitamin B6 presented only a direct effect, and folate presented only an indirect effect through methylation (Table 2). Neopterin showed a significant risk effect via vitamin B6 and KTR showed a significant risk effect only on the log2 scale ($P<0.01$).

DISCUSSION

Several factors of the OCM were recently found to be associated with lung cancer within the EPIC prospective cohort (6). In this nested case-control study of lung cancer, we applied SEM on the same dataset used in the original study to model a set

of pathways between factors involved in the OCM in relation to the risk of lung cancer.

We modelled the OCM using several linear equations, each describing relations between selected biomarkers and lung cancer risk. The methylation part of the OCM pathway was represented by methionine which showed strong inverse association with risk, thus suggesting DNA methylation as a potential aetiological mechanism in lung carcinogenesis. Vitamin B6 and folate, here representing inflammation and nucleotide synthesis, respectively, were also associated to risk, but these associations were weaker and the statistical significances were conditional on the variable transformation, i.e. vitamin B6 was significant in the log2 scale but not in the SD scale, meanwhile folate was significant in the SD scale, but not in the log2. In their respective pathways, methionine was dependent on homocysteine, vitamin B12, vitamin B6, and indirectly by serine (via homocysteine), while folate was dependent on vitamin B2 (and serine in the log2 scale). The transsulfuration branch of the OCM, here represented by cysteine as a final intermediate marker, was not associated with lung cancer risk. In addition, according to AIC, model A fitted considerably better than the others, for both scales.

In a case-control study nested with the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) study, Hartman et al (3) evaluated the associations between serum vitamins B6, B12, folate and homocysteine with lung cancer. In this study, vitamin B6 was found to be inversely associated with lung cancer, an association recently strongly replicated by Johansson et al. (6) using the same dataset as the current study. While the results of Hartman et al (3) are difficult to interpret because of the limited sample size and absence of never smokers, Johansson et al.

confirmed the association in never smokers, and described a seemingly independent, and equally strong, inverse association of methionine with lung cancer risk. Johansson et al. also reported an inverse association of folate with risk, but this observation was confined to former and current smokers. In model B (Figure 2), vitamin B6 exerted an indirect effect on risk by being associated with methionine levels. Vitamin B6 acts as a co-factor in more than 100 mechanisms in the body, and it is possible that the strong effect of vitamin B6 noted by Hartman et al. and Johansson et al. reflects mechanisms independent of the OCM and inflammation as suggested by Lee et al (24).

Our SEM models also quantify the associations between several intermediate variables involved in the OCM, including the re-methylation loop in which homocysteine and methionine are associated with each other. The reverse loop expresses two relations: 1) decreasing methionine, possibly used for SAM formation, results in increasing homocysteine,; and 2) increasing homocysteine results in increasing methionine, with vitamins B6 and B12 as cofactors. Starting from model C where both the methylation and nucleotide synthesis pathways were introduced, as well as the transsulfuration pathway, vitamin B6 was hypothesised to affect risk via its association with methionine (Table 2), as well as directly through inflammation and immune activation on the log2 scale (Model D).

To our knowledge, this is the first study which has attempted to evaluate the role and the relative contributions of the multiple components of the OCM in relation to lung cancer. The study has a prospective design, in which serum levels were measured several years before the disease onset. The main limitation of our study relates to the study design in which blood samples were drawn on one occasion for

each participant, thus preventing us from investigating time-related relations among the investigated factors. Theoretical models were tested with the consideration that they were confirmed when the observed linear correlations agreed with the hypothesized. It was not possible to include all biomarkers involved in the mechanism with a lack of information about enzymes and gene methylation. In addition to the possible misspecification, it is also possible that some polymorphism related to the enzymes involved in the pathways has an influence on the relations that were found.

The effect of methylation and its components is evident on both scales (SD and log2), though the nucleotide synthesis and inflammation pathways showed some differences between the two scales.

In conclusion, our data primarily support a role of DNA methylation in lung cancer aetiology. Factors representing nucleotide synthesis, inflammation and immune activation also showed some association with risk, but these associations were weaker and less consistent.

REFERENCES

- (1) World Health Organization. "Cancer". Fact sheet N°297. February 2009. Switzerland. (<http://www.who.int/mediacentre/factsheets/fs297/en/>). (Accessed August 28, 2010)
- (2) Ames BN. Low micronutrient intake may accelerate the degenerative diseases of aging through allocation of scarce micronutrients by triage. *PNAS*. 2006; **103**(47):17589-17594.
- (3) Hartman TJ, Woodson K, Stolzenberg-Solomon R, Virtamo J, Selhub J, Barrett MJ et al. Association of the b-vitamins pyridoxal 5'-phosphate (b6), B12, and folate with lung cancer risk in older men. *Am J Epidemiol*. 2001; **153**(7):688-694.
- (4) Ulrich CM, Reed MC, Hijhout HF. Modeling folate, one-carbon metabolism, and DNA methylation. *Nutrition Reviews*. 2008; **66**(1): S27-S30.
- (5) Ulrich CM, Neuhouser M, Liu AY, Boynton A, Gregory JF 3rd, Shane B, et al. Mathematical modelling of folate metabolism: predicted effects of genetic polymorphisms on mechanisms and biomarkers relevant to carcinogenesis. *Cancer Epidemiol Biomarkers Prev*. 2008; **17**(7):1822-1831.
- (6) Johansson M, Relton C, Ueland PM, Stein Emil Vollset SE, Midttun Ø, Nygard O et al. Serum B vitamin levels and risk of lung cancer. *JAMA*. 2010;**303**(23):2377-2385.
- (7) Robertson CE, Vineis P. Biomarkers for Nutrient-Gene Interactions. In: Choi S-W, Friso S, eds. *Nutrient-Gene Interactions in Cancer*. Boca Raton, FL: CRC Press; 2006: 37-55.
- (8) Choi S-W, Friso S. Interaction between folate and methylene-tetrahydrofolate reductase gene in cancer. In: Choi S-W, Friso S. eds. *Nutrient-Gene Interactions in Cancer*. Boca Raton, FL: CRC Press; 2006: 57-85.
- (9) Das PM, Signal R. DNA methylation and cancer. *J Clin Oncol*. 2004; **22**:4632-4642.
- (10) Selhub J. Homocysteine metabolism. *Annu. Rev Nutr*. 1999; **19**:217-246.
- (11) Vaissière T, Cuenin C, Paliwal A, Vineis P, Hoek G, Krzyzanowski M et al. Quantitative analysis of DNA methylation after whole bisulfite amplification of a minute amount of DNA from body fluids. *Epigenetics*. 2009; **4**(4):221-230.

- (12) Paliwal A, Vaissière T, Herczeg Z. Quantitative detection of DNA methylation states in minute amounts of DNA from body fluids. *Methods*. 2010; **52** (3): 242-247.
- (13) Leklen JE. Vitamin B6. In: Ziegler EE, Filer LJ Jr, eds. *Present Knowledge in nutrition*. Washington,DC: ILSI Press; 1996: 174-183.
- (14) Davis SR, Quinlivan EP, Stacpoole PW, Gregory JF 3rd. Plasma glutathione and cystathionine concentrations are elevated but cysteine flux is unchanged by dietary vitamin B-6 restriction in young men and women. *J Nutr*. 2006; 136:373-378.
- (15) Price JM, Brown RR, Yess N. Testing the functional capacity of the tryptophan-niacin pathway in man by analysis of urinary metabolites. *Adv Metab disord*. 1965; **2**:159-225.
- (16) Seccareccia F, Zuccaro P, Pacifici R, Meli P, Pannozzo F, Freeman KM et al. Serum cotinine as a marker of environmental tobacco smoke exposure in epidemiological studies: the experience of the MATISS Project. *Eur J Epidemiol*. 2003; **18**(6):487-492.
- (17) Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M et al. European Prospective Investigation Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002; **5**:1113-1124.
- (18) Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P. Estimation of health effects of prenatal methylmercury exposure using structural equation models. *Environ Health*. 2002; **1**(1):2.
- (19) Scott JM. Folate and vitamin B12. *Proc Nutr Soc*. 1999; **58**(2):441-448.
- (20) Johansson M, Van Guelpen B, Vollset SE, Hulthén J, Bergh A, Key T et al. One-carbon metabolism and prostate cancer risk: prospective investigation of seven circulating B vitamins and metabolites. *Cancer Epidemiol Biomarkers Prev*. 2009 ; **18**:1538-1543.
- (21) Muthén BO, Satorra A. Complex sample data in Structural Equation Modeling. *Sociol Methodol*, 1995, **25**: 267-316.
- (22) Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén; 1998-2010.

- (23) Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Chapman & Hill/CRC; Washington, DC. 2004.
- (24) Lee JE, Li H, Giovannucci E, Lee IM, Selhub J, Stampfer M et al. Prospective study of plasma vitamin B6 and risk of colorectal cancer in men. *Cancer Epidemiol Biomarkers Prev.* 2009; **18**(4):1197-1202.

Figure 1. Diagram of the pathways of one-carbon metabolism. Labile methyl groups are supplied by dietary serine, choline (via betaine) and methionine. One-carbon units are derived from serine through the activity of the vitamin B₆-dependent serine hydroxymethyltransferase, which generates 5,10-methylenetetrahydrofolate (5,10-methyleneTHF). 5,10-methyleneTHF in turn is then reduced to 5-methyltetrahydrofolate (5-methylTHF) - the predominant form of folate in the circulation - in an irreversible reaction catalyzed by vitamin B₂-dependent methylenetetrahydrofolate reductase (MTHFR). 5-methylTHF serves as a methyl donor in a reaction converting homocysteine to methionine, where vitamin B₁₂ serves a co-factor. Homocysteine can also be metabolized to cysteine through the sequential action of two vitamin B₆-dependent enzymes, i.e. the transsulfuration pathway. The methionine derivative, S-adenosylmethionine (SAM), is the universal methyl donor for the methylation of a vast variety of molecules, including DNA (20).

Figure 2. Complete pathway: each colour represents one branch of the SEM (age and sex are not included in the figure but they are included in the equations as following).

Equations:

Model A: The methylation pathway.

1. lung cancer = cotinine + methionine
2. methionine = cotinine+sex+age+homocysteine +vitamin B12+folate+vitamin B6
3. homocysteine = cotinine + sex + age + serine + methionine

Model B: The methylation and the nucleotide synthesis pathways.

1. lung cancer =cotinine+ methionine +folate
2. methionine = cotinine+sex+age+homocysteine + vitamin B12+folate+vitamin B6
3. homocysteine = cotinine+sex+age+ serine + methionine
4. folate = cotinine+sex+age+serine + vitamin B2

Model C: The methylation, the nucleotide synthesis and the transsulfuration pathways.

1. lung cancer = cotinine+methionine +folate+ cysteine+vitamin B6+KTR+neopterin
2. methionine = cotinine+sex+age+homocysteine + vitamin B12+folate+vitamin B6
3. homocysteine = cotinine+sex+age+serine+methionine
4. folate = cotinine+sex+age+ serine + vitamin B2
5. cysteine = cotinine+sex+age+vitamin B6+cystathionine

6. cystathionine= cotinine+sex+age+vitamin B6+ homocysteine+serine

Model D: The methylation, the nucleotide synthesis, the transsulfuration and the inflammation and immune activation pathways.

1. lung cancer = cotinine+methionine +folate+ cysteine+vitamin B6+KTR+neopterin
2. methionine = cotinine+sex+age+homocysteine + vitamin B12+folate+vitamin B6
3. homocysteine = cotinine+sex+age+serine+methionine
4. folate = cotinine+sex+age+serine+vitamin B2
5. cysteine = cotinine+sex+age+vitamin B6+cystathionine
6. cystathionine= cotinine+sex+age+vitamin B6+homocysteine+serine
7. vitamin B6= cotinine+sex+age+KTR + neopterin

Model E: The methylation, the nucleotide synthesis, and the inflammation and immune activation pathways

1. lung cancer = cotinine+methionine +folate+ cysteine+vitamin B6
2. methionine = cotinine+sex+age+homocysteine+vitamin B12+folate+vitamin B6
3. homocysteine = cotinine+sex+age+serine+methionine
4. folate = cotinine+sex+age+serine+vitamin B2
7. vitamin B6= cotinine+sex+age+KTR+neopterin

Table 1: Summary of Plasma Concentrations of B Vitamins and Metabolites by Lung Cancer Status (891 Cases / 1747 Controls)

Variable	Mean (SD)			
	Cases (891)		Control (1747)	
Homocysteine	13.89	(6.79)	13.03	(4.45)
Methionine	28.57	(7.22)	30.44	(7.57)
Folate	15.05	(9.63)	18.25	(15.75)
Vitamin B12	380.54	(235.57)	371.12	(215.80)
Vitamin B6	42.49	(50.74)	52.28	(51.48)
Vitamin B2	25.36	(36.47)	28.92	(49.01)
Serine	146.65	(36.26)	148.40	(35.36)
Cysteine	295.45	(34.75)	299.70	(35.85)
Cystathionine	0.23	(0.15)	0.25	(0.34)
Neopterin	18.41	(6.84)	18.36	(7.09)
KTR	0.02	(0.01)	0.02	(0.01)

Abbreviations: SD, standard deviation.

Table 2: SEMs results for Models A, B, C, D and E for both scales SD and base 2 logarithm

		Model A				Model B				Model C				Model D				Model E			
1. Lung Cancer status	scale	OR/β	CI 95%		P	OR/β	CI 95%		P	OR/β	CI 95%		P	OR/β	CI 95%		P	OR/β	CI 95%		P
Cotinine	SD	2.60	2.35	2.88	<0,01	2.55	2.30	2.82	<0,01	2.57	2.32	2.85	<0,01	2.60	2.34	2.89	<0,01	2.54	2.29	2.81	<0,01
	SD	2.60	2.35	2.88	<0,01	2.53	2.29	2.81	<0,01	2.55	2.3	2.829	<0,01	2.53	2.27	2.82	0.00	2.47	2.23	2.74	<0,01
Methionine	SD	0.81	0.73	0.90	<0,01	0.81	0.73	0.90	<0,01	0.81	0.73	0.89	<0,01	0.82	0.74	0.91	<0,01	0.81	0.73	0.90	<0,01
	log2	0.52	0.39	0.69	<0,01	0.52	0.39	0.69	<0,01	0.52	0.39	0.688	<0,01	0.57	0.43	0.76	0.00	0.55	0.41	0.73	<0,01
Folate	SD					0.87	0.78	0.98	0.02	0.87	0.77	0.97	0.01	0.88	0.79	0.99	0.04	0.88	0.78	0.99	0.03
	log2					0.88	0.79	0.99	0.03	0.88	0.782	0.983	0.02	0.94	0.83	1.06	0.29	0.93	0.83	1.05	0.26
Cysteine	SD									1.05	0.96	1.15	0.30	1.03	0.94	1.14	0.51				
	log2									1.35	0.768	2.354	0.30	1.22	0.69	2.15	0.50				
Vitamin B6	SD													0.97	0.85	1.09	0.57	0.96	0.85	1.09	0.52
	log2													0.85	0.75	0.96	0.01	0.84	0.74	0.95	0.01
KTR	SD													1.08	0.97	1.21	0.14				
	log2													1.13	0.89	1.43	0.31				
Neopterin	SD													1.03	0.93	1.15	0.53				
	log2													1.09	0.89	1.33	0.40				
2. Methionine		β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P
Cotinine	SD	-0.47	-0.64	-0.29	<0,01	-0.47	-0.64	-0.29	<0,01	-0.47	-0.64	-0.29	<0,01	-0.47	-0.64	-0.29	<0,01	-0.47	-0.64	-0.29	<0,01
	SD	-0.13	-0.18	-0.08	<0,01	-0.13	-0.18	-0.08	<0,01	-0.13	-0.18	-0.08	<0,01	-0.13	-0.18	-0.08	<0,01	-0.13	-0.18	-0.08	<0,01
Sex	SD	0.38	0.16	0.60	<0,01	0.38	0.16	0.60	<0,01	0.38	0.16	0.60	<0,01	0.38	0.16	0.60	<0,01	0.38	0.16	0.60	<0,01
	log2	0.22	0.13	0.30	<0,01	0.22	0.13	0.30	<0,01	0.22	0.13	0.30	<0,01	0.22	0.13	0.30	<0,01	0.22	0.13	0.30	<0,01
Age	SD	-0.05	-0.07	-0.04	<0,01	-0.05	-0.07	-0.04	<0,01	-0.05	-0.07	-0.04	<0,01	-0.05	-0.07	-0.04	<0,01	-0.05	-0.07	-0.04	<0,01
	log2	-0.03	-0.03	-0.02	<0,01	-0.03	-0.03	-0.02	<0,01	-0.03	-0.03	-0.02	<0,01	-0.03	-0.03	-0.02	<0,01	-0.03	-0.03	-0.02	<0,01
Homocysteine	SD	2.53	2.00	3.06	<0,01	2.53	2.00	3.06	<0,01	2.53	2.00	3.06	<0,01	2.53	2.00	3.06	<0,01	2.53	2.00	3.06	<0,01
	log2	2.22	1.82	2.62	<0,01	2.22	1.82	2.62	<0,01	2.22	1.82	2.62	<0,01	2.22	1.82	2.62	<0,01	2.22	1.82	2.62	<0,01
Cobalamin	SD	0.32	0.18	0.46	<0,01	0.32	0.18	0.46	<0,01	0.32	0.18	0.46	<0,01	0.32	0.18	0.46	<0,01	0.32	0.18	0.46	<0,01
	log2	0.40	0.30	0.50	<0,01	0.40	0.30	0.50	<0,01	0.40	0.30	0.50	<0,01	0.40	0.30	0.50	<0,01	0.40	0.30	0.50	<0,01
Folate	SD	0.25	-0.14	0.64	0.21	0.25	-0.14	0.64	0.21	0.25	-0.14	0.64	0.21	0.25	-0.14	0.64	0.21	0.25	-0.14	0.64	0.21
	log2	0.34	0.24	0.43	<0,01	0.34	0.24	0.43	<0,01	0.34	0.24	0.43	<0,01	0.34	0.24	0.43	<0,01	0.34	0.24	0.43	<0,01
Vitamin B6	SD	0.10	0.00	0.19	0.05	0.10	0.00	0.19	0.05	0.10	0.00	0.19	0.05	0.10	0.00	0.19	0.05	0.10	0.00	0.19	0.05
	log2	0.01	-0.03	0.06	0.55	0.01	-0.03	0.06	0.55	0.01	-0.03	0.06	0.55	0.01	-0.03	0.06	0.55	0.01	-0.03	0.06	0.55
3. Homocysteine		β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P

		Model A				Model B				Model C				Model D				Model E			
Cotinine	SD	-0.13	-0.27	0.01	0.07	-0.13	-0.27	0.01	0.07	-0.13	-0.27	0.01	0.07	-0.13	-0.27	0.01	0.07	-0.13	-0.27	0.01	0.07
	SD	-0.07	-0.12	-0.02	0.01	-0.07	-0.12	-0.02	0.01	-0.07	-0.12	-0.015	0.01	-0.07	-0.12	-0.02	0.01	-0.07	-0.12	-0.02	0.01
Sex	SD	-1.60	-2.23	-0.97	<0,01	-1.60	-2.23	-0.97	<0,01	-1.60	-2.23	-0.97	<0,01	-1.60	-2.23	-0.97	<0,01	-1.60	-2.23	-0.97	<0,01
	log2	-0.81	-0.97	-0.66	<0,01	-0.81	-0.97	-0.66	0.00	-0.81	-0.97	-0.66	0.00	-0.81	-0.97	-0.66	<0,01	-0.81	-0.97	-0.66	<0,01
Age	SD	0.00	-0.02	0.01	0.76	0.00	-0.02	0.01	0.76	-0.002	-0.02	0.01	0.76	0.00	-0.02	0.01	0.76	0.00	-0.02	0.01	0.76
	log2	0.00	-0.01	0.01	0.84	0.00	-0.01	0.01	0.84	0.001	-0.005	0.01	0.84	0.00	-0.01	0.01	0.84	0.00	-0.01	0.01	0.84
Serine	SD	1.72	1.01	2.43	<0,01	1.72	1.01	2.43	<0,01	1.72	1.01	2.43	<0,01	1.72	1.01	2.43	<0,01	1.72	1.01	2.43	<0,01
	log2	2.35	1.94	2.77	<0,01	2.35	1.94	2.77	<0,01	2.35	1.94	2.77	0.00	2.35	1.94	2.77	<0,01	2.35	1.94	2.77	<0,01
Methionine	SD	-2.76	-4.06	-1.45	<0,01	-2.76	-4.06	-1.45	<0,01	-2.76	-4.06	-1.45	<0,01	-2.76	-4.06	-1.45	<0,01	-2.76	-4.06	-1.45	<0,01
	log2	-3.75	-4.50	-2.99	<0,01	-3.75	-4.50	-2.99	<0,01	-3.75	-4.50	-2.99	0.00	-3.75	-4.50	-2.99	<0,01	-3.75	-4.50	-2.99	<0,01
4. Folate		β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P
Cotinine	SD					-0.14	-0.17	-0.11	<0,01	-0.14	-0.17	-0.11	<0,01	-0.14	-0.17	-0.11	<0,01	-0.14	-0.17	-0.11	<0,01
	SD					-0.17	-0.20	-0.14	<0,01	-0.17	-0.20	-0.14	<0,01	-0.17	-0.20	-0.14	0.00	-0.17	-0.20	-0.14	<0,01
Sex	SD					0.07	-0.02	0.15	0.11	0.07	-0.02	0.15	0.11	0.07	-0.02	0.15	0.11	0.07	-0.02	0.15	0.11
	log2					-0.03	-0.09	0.04	0.42	-0.03	-0.09	0.04	0.42	-0.03	-0.09	0.04	0.42	-0.03	-0.09	0.04	0.42
Age	SD					0.01	0.00	0.01	<0,01	0.01	0.00	0.01	<0,01	0.01	0.00	0.01	<0,01	0.01	0.00	0.01	<0,01
	log2					0.01	0.00	0.01	<0,01	0.01	0.00	0.01	<0,01	0.01	0.00	0.01	0.00	0.01	0.00	0.01	<0,01
Serine	SD					-0.03	-0.10	0.03	0.33	-0.03	-0.10	0.03	0.33	-0.03	-0.10	0.03	0.33	-0.03	-0.10	0.03	0.33
	log2					-0.19	-0.28	-0.09	<0,01	-0.19	-0.28	-0.09	<0,01	-0.19	-0.28	-0.09	0.00	-0.19	-0.28	-0.09	<0,01
Vitamin B2	SD					0.12	0.01	0.22	0.03	0.12	0.01	0.22	0.03	0.12	0.01	0.22	0.03	0.12	0.01	0.22	0.03
	log2					0.20	0.17	0.23	<0,01	0.20	0.17	0.23	<0,01	0.20	0.17	0.23	0.00	0.20	0.17	0.23	<0,01
5. Cystathionine		β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P	β	CI 95%		P
Cotinine	SD									-0.05	-0.08	-0.02	0.00	-0.05	-0.08	-0.02	<0,01				
	log2									-0.05	-0.08	-0.03	0.00	-0.05	-0.08	-0.03	<0,01				
Sex	SD									0.00	-0.07	0.08	0.97	0.00	-0.07	0.08	0.97				
	log2									-0.04	-0.09	0.02	0.20	-0.04	-0.09	0.02	0.20				
Age	SD									0.01	0.00	0.01	0.18	0.01	0.00	0.01	0.18				
	log2									0.02	0.01	0.02	0.00	0.02	0.01	0.02	<0,01				
Homocysteine	SD									0.06	0.03	0.10	<0,01	0.06	0.03	0.10	<0,01				
	log2									0.33	0.27	0.40	0.00	0.33	0.27	0.40	<0,01				
Serine	SD									0.05	0.01	0.08	0.01	0.05	0.01	0.08	0.01				

		Model A			Model B			Model C				Model D				Model E			
Vitamin B6	log2							0.09	0.01	0.18	0.04	0.09	0.01	0.18	0.04				
	SD							-0.02	-0.04	-0.01	0.00	-0.02	-0.04	-0.01	<0,01				
	log2							-0.03	-0.06	-0.01	0.01	-0.03	-0.06	-0.01	0.01				
6.Cysteine		β	CI 95%	P	β	CI 95%	P	β	CI 95%	P		β	CI 95%	P		β	CI 95%	P	
Cotinine	SD							-0.08	-0.12	-0.05	<0,01	-0.08	-0.12	-0.05	<0,01				
	log2							-0.01	-0.02	-0.01	<0.01	-0.01	-0.02	-0.01	0.00				
Sex	SD							-0.11	-0.17	-0.04	<0.01	-0.11	-0.17	-0.04	<0,01				
	log2							-0.02	-0.03	-0.01	<0.01	-0.02	-0.03	-0.01	0.00				
Age	SD							0.05	0.04	0.05	<0,01	0.05	0.04	0.05	<0,01				
	log2							0.01	0.01	0.01	<0.01	0.01	0.01	0.01	0.00				
Cystathionine	SD							0.04	0.00	0.08	0.04	0.04	0.00	0.08	0.04				
	log2							0.03	0.02	0.04	<0.01	0.03	0.02	0.04	0.00				
Vitamin B6	SD							0.05	0.01	0.10	0.02	0.05	0.01	0.10	0.02				
	log2							0.01	0.01	0.02	<0.01	0.01	0.01	0.02	0.00				
7. Vitamin B6		β	CI 95%	P	β	CI 95%	P	β	CI 95%	P		β	CI 95%	P		β	CI 95%	P	
Cotinine	SD											-0.16	-0.19	-0.12	<0,01	-0.16	-0.19	-0.12	<0,01
	log2											-0.26	-0.30	-0.23	<0.01	-0.26	-0.30	-0.23	<0.01
Sex	SD											0.04	-0.04	0.12	0.34	0.04	-0.04	0.12	0.34
	log2											-0.04	-0.11	0.03	0.24	-0.04	-0.11	0.03	0.24
Age	SD											0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.95
	log2											0.00	-0.01	0.00	0.05	0.00	-0.01	0.00	0.05
KTR	SD											-0.03	-0.07	0.01	0.12	-0.03	-0.07	0.01	0.12
	log2											-0.13	-0.21	-0.05	<0.01	-0.13	-0.21	-0.05	<0.01
Neopterin	SD											-0.07	-0.11	-0.04	<0,01	-0.07	-0.11	-0.04	<0,01
	log2											-0.18	-0.25	-0.11	<0.01	-0.18	-0.25	-0.11	<0.01
AIC	SD		19970.76			27136.29			33822.93				48128.87				34551.56		
	log2		9469.50			15594.46			18512.85				25177.66				22258.10		

* OR obtained from conditional logistic regression within structural equation model (Equations 1); β obtained from linear regression within structural equation model.

Abbreviations: SD, standard deviation; log2, base 2 logarithm; OR, odds ratio; β, linear regression coefficient; CI, confidence interval; P, P value; KTR, kynurenin/tryptophan ratio; AIC, Akaike information criterion.

4.3. MANUSCRITO 3

Artigo não submetido. Sugestão de periódico: *Epidemiology*.

Latent model for DNA methylation, nucleotide synthesis and immune activation for lung cancer risk

Valéria Troncoso Baltar^{1,8}, Wei Xun², Shu-Chun Chuang², Caroline Relton³, Per Magne Ueland^{4,5}, Stein Emil Vollset^{6,7}, Øivind Midttun⁸, Pietro Ferrari⁹, Mattias Johansson⁹, Nadia Slimani⁹, Mazda Jenab⁹, Françoise Clavel-Chapelon^{10,11,12}, Marie-Christine Boutron-Ruault^{10,11,12}, Guy Fagherazzi^{10,11,12}, Rudolf Kaaks¹³, Sabine Rohrmann¹⁴, Heiner Boeing¹⁵, Cornelia Weikert¹⁵, H. Bas Bueno-de-Mesquita¹⁶, Hendriek Boshuizen¹⁶, Carla H. van Gils¹⁷, N Charlotte Onland-Moret¹⁷, Antonio Agudo¹⁸, Aurelio Barricarte^{19,23}, Carmen Navarro^{20,23}, Laudina Rodríguez²¹, José Maria Huerta Castaño^{22,23}, Nerea Larrañaga^{23,24}, Kay-Tee Khaw²⁵, Nick Wareham²⁶, Naomi E. Allen²⁷, Francesca Crowe²⁷, Valentina Gallo², Teresa Norat², Vittorio Krogh²⁸, Giovanna Masala²⁹, Salvatore Panico³⁰, Carlota Sacerdote³¹, Rosario Tumino³², Antonia Trichopoulou^{33,34}, Pagona Lagiou^{33,35}, Dimitrios Trichopoulos^{35,36}, Torgny Rasmussen³⁷, Göran Hallmans³⁸, Nina Roswall³⁹, Anne Tjønneland³⁹, Elio Riboli², Paul Brennan⁹, Paolo Vineis^{31,40}

¹Department of Epidemiology, Faculty of Public Health, University of São Paulo, Av. Dr Arnaldo, 715. CEP 01246-904. São Paulo, São Paulo, Brazil

²Department of Epidemiology & Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, London, UK

³Institute for Ageing and Health, Newcastle University, Campus for Ageing and Vitality, Newcastle, UK

⁴Section for Pharmacology, Institute of Medicine, University of Bergen, Bergen, Norway

⁵Haukeland University Hospital, Bergen, Norway

⁶Department of Public Health and Primary Health Care, University of Bergen, Bergen, Norway

⁷Division of Epidemiology, Norwegian Institute of Public Health, Bergen, Norway

⁸Bevital AS, Bergen, Norway

⁹International Agency for Research on Cancer, Lyon, France

¹⁰Institut National de la Santé et de la Recherche Médicale, INSERM U780-IFR169, Villejuif, France

¹¹Institut Gustave Roussy, Villejuif, France

¹²Paris South University, Villejuif, France

¹³Division of Cancer Epidemiology, German Cancer Research Centre, Heidelberg, Germany

- ¹⁴Division of Clinical Epidemiology, German Cancer Research Centre, Heidelberg, Germany
- ¹⁵Department of Epidemiology, German Institute of Human Nutrition, Nuthetal, Germany
- ¹⁶National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands
- ¹⁷Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
- ¹⁸Unit of Nutrition, Environment and Cancer, Catalan Institute of Oncology, L'Hospitalet de Llobregat, Spain
- ¹⁹Epidemiology, Prevention and Promotion Health Service, Public Health Institute of Navarra, Navarra, Spain
- ²⁰Department of Epidemiology, Regional Council of Health and Consumer Affairs, Murcia, Spain
- ²¹Public Health and Participation Directorate, Health and Health Care Services Council, Oviedo, Asturias, Spain
- ²²Department of Epidemiology, Murcia Regional Health Authority, Murcia, Spain
- ²³CIBER of Epidemiology and Public Health, Barcelona, Spain
- ²⁴Epidemiology Unit, Public Health Department of Gipuzkoa, Basque Government San Sebastián, Spain
- ²⁵Cancer registry of anada, Andalusian School of Public Health, Granada, Spain
- ²⁶Clinical Gerontology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
- ²⁷MRC Epidemiology Unit, Strangeways Research Laboratory, Strangeways Research Laboratory, Cambridge, UK
- ²⁸Cancer Epidemiology Unit, University of Oxford, Oxford, UK
- ²⁹Environmental Epidemiology and Cancer Registry Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italia
- ³⁰Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Centre, Scientific Institute of Tuscany, Florence, Italy
- ³¹Department of Clinical and Experimental Medicine, Federico II University of Naples, Naples, Italy
- ³²Human Genetic Foundation (HuGeF), Torino Italy,
- ³³Ragusa Cancer Registry, Azienda Ospedaliera "Civile M.P. Arezzo", Ragusa, Italy
- ³⁴WHO Collaborating Center for Food and Nutrition Policies, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece
- ³⁵Hellenic Health Foundation, Athens, Greece
- ³⁶Department of Epidemiology, Harvard School of Public Health, Harvard University, Boston, MA. USA
- ³⁷Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden
- ³⁸Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden
- ³⁹Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark
- ⁴⁰Department of Epidemiology & Public Health, MRC-HPA Centre for Environment and Health, Imperial College, London, UK
- [§]Corresponding author

Email addresses: VTB: vbaltar@usp.br, WX: wei.xun06@imperial.ac.uk, SCC: s-c.chuang@imperial.ac.uk, CR: c.l.relton@newcastle.ac.uk, PMU: per.ueland@ikb.uib.no, SEV: Vollset@uib.no, ØM: bjorn.midttun@farm.uib.no, PF: pietro2204@yahoo.com, MJ: johanssonm@fellows.iarc.fr, NS: slimani@iarc.fr, MJ: jenab@iarc.fr, FC-C: clavel@igr.fr, M-CB-R: boutron@igr.fr, GF: guy.fagherazzi@igr.fr, RK: r.kaaks@dkfz-heidelberg.de, SR: sabine.rohrmann@ifspm.uzh.ch, HB: boeing@mail.dife.de, CW: weikert@dife.de, HBBd-M: Bas.Bueno.De.Mesquita@rivm.nl, HCB: Hendriek.Boshuizen@rivm.nl, CHvG: C.vanGils@umcutrecht.nl, PP: P.H.M.Peeters@umcutrecht.nl, AA: a.agudo@iconcologia.net, AB: abarricg@cfnavarra.es, CN: carmen.navarro@carm.es, LR: laudina.rodriquezsuarez@asturias.org, JMHC: jmhuerta.carm@gmail.com, NL: epidem3-san@ej-gv.es, MJSP: mariajose.sanchez.easp@juntadeandalucia.es, K-TK: kk101@medschl.cam.ac.uk, NW: nick.wareham@mrc-epid.cam.ac.uk, NA: Naomi.allen@ceu.ox.ac.uk, FC: francesca.crowe@ceu.ox.ac.uk, VG: v.gallo@imperial.ac.uk, TN: t.norat@imperial.ac.uk, GT: giovanna.tagliabue@istitutotumori.mi.it, GM: g.masala@ispo.toscana.it, SP: spanico@unina.it, CS: carlotta.sacerdote@cpo.it, RT: rtumino@tin.it, AT: atrachopoulou@hhf-greece.gr, PL: pdlagiou@med.uoa.gr, CB: cbamia@nut.uoa.gr, TR: torgny.rasmuson@onkologi.umu.se, GH: goran.hallmans@nutrires.umu.se, NR: roswall@cancer.dk, AT: annet@cancer.dk, ER: e.riboli@imperial.ac.uk, PB: brennan@iarc.fr, PV: p.vineis@imperial.ac.uk

Short running title: Latent model for lung cancer risk

Abbreviations: LC: lung cancer; OCM: one-carbon metabolism; SEM- Structural equation model; EPIC: European Prospective Investigation into Cancer and Nutrition; PLP: vitamin B₆; Fol: folate; Rib: vitamin B₂; Ser: serine; Gly: glycine; Npt: neopterin; KTR: kynurenin/thriptophan ratio; Cot: cotinine.

ABSTRACT

Lung cancer (LC) continues to be the most common cancer death in the world. In addition to tobacco exposure, low-micronutrient intake has been linked to LC. The intake of micronutrients is the main source of vitamins and amino acids involved in the one-carbon metabolism (OCM) which is considered key in maintaining DNA integrity, regulating gene expression, and may thus affect carcinogenesis. Two important branches of the OCM are related to cancer: DNA methylation and nucleotide synthesis. Immune activation is involved in the aging process in normal healthy individuals as well as in a number of pathologies, including cancer. In order to investigate how these three pathways are related to LC we applied structural equation models (SEM) to estimate three latent variables in the pathway to LC in addition to tobacco exposure (measured by cotinine levels). Each latent variable represents one of the mechanisms: DNA methylation, DNA synthesis and immune activation. We have conducted a study using the dataset from a case-control nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. We found a direct and protective effect for DNA methylation ($p=0.018$) and Immune activation ($p=0.037$), whereas nucleotide synthesis presented a direct risk effect ($p=0.041$) and an indirect protective effect ($p=0.013$). In conclusion, our results support a role of DNA methylation and immune activation in LC aetiology. Tobacco effect remains as the factor with the highest impact in lung cancer ($p<0.001$).

Keywords: lung cancer, methylation, nucleotide synthesis, immune activation, structural equation model, latent model

INTRODUCTION

Lung cancer (LC) continues to be the most common cancer death in the world. In the GLOBOCAN project (<http://globocan.iarc.fr/factsheets/cancers/lung.asp>), in 2008 there were 1.61 million new cases of LC, representing 12.7% of all new cancer cases. In this year 1.38 million deaths (18.2% of the total of cancer deaths) were due to LC. The reduction of tobacco consumption remains the most appropriate strategy to reduce incidence (Brennan et al 2010). In addition to exposures to carcinogens or carcinogenic mixtures such as smoking and air pollution, low-micronutrient intake has been linked to degenerative diseases including LC (Ames 2006, Hartman et al 2001). The intake of micronutrients is the main source of vitamins and amino acid involved in the one-carbon metabolism which is considered key in maintaining DNA integrity, regulating gene expression, and may thus affect carcinogenesis. Two important branches of the one-carbon metabolism are related to cancer: DNA methylation and nucleotide synthesis.

DNA methylation is the addition of a methyl group to a protein (such as histones) in the DNA nucleus (Stover 2009). Disturbance in DNA methylation occurs during carcinogenic processes (Ulrich 2008). Gene hypermethylation might make the gene inaccessible, resulting in a silencing of the gene, and this can be a tumour suppressor gene (Das and Signal 2004). On the other hand, global DNA methylation is generally lower in tumour cells when compared to normal cells (Feinberg and Vogelstein 1983, Lengauer 2003). Genome-wide hypomethylation

has been linked to chromosomal instability (Eden 2003) promoting tumourgenesis (Lednigauer 2003).

The essential conversion for the synthesis and repair of DNA is the methylation of deoxyuridylate (dUMP) to form thymidylate (dTMP). Folate is ingested from food and occurs in the mitochondria in the form of tetrahydrofolate (THF). THF is condensed with the amino acid serine (vitamin B6 dependent) to generate 5,10-methylene-THF (5,10MeTHF). 5,10MeTHF is essential to dTMP biosynthesis (Stover 2009, Fenech 2010). When the dUMP/dTMP ratio increases, Uracil accumulates in the DNA instead of thymine. In most cells Uracil remains as a mutagenic lesion and must be repaired (Visnes et al. 2009).

These two branches of the one-carbon metabolism are linked to each other by the dependence of folate as a methyl transfer. The 5,10MeTHF used in dTMP biosynthesis is reduced to 5MeTHF in an irreversible reaction catalyzed by vitamin B2 which is dependent on methylenetetrahydrofolate reductase (MTHFR). 5MeTHF serves as a methyl donor to convert homocysteine to methionine, where vitamin B12 serves as co-factor. The methionine derivative, S-adenosylmethionine (SAM), is the universal methyl donor for the methylation of a vast variety of molecules, including DNA (Das and Signal, 2004, Stover, 2009, Scott, 1999).

Another branch of the OCM pathway is the transsulfuration (Davis et al. 2006, Gregory 3rd et al. 2000). We did not hypothesize that this branch is related to LC directly. Because serine is the one carbon donor to homocysteine in the transsulfuration branch, we included it in the methylation pathway (control of homocysteine levels).

Immune activation is another important mechanism for cancer prevention. It is involved in the aging process in normal healthy individuals as well as in a number of pathologies, including cancer (Spencer et al. 2010). Neopterin is a metabolite of GTP (guanosine triphosphate) and it is considered as a biomarker of macrophage activation and has been used clinically for bacterial and viral infection assessment, autoimmune diseases and malignant conditions (Spencer et al. 2010). Additionally, thryptophan degradation represents an antimicrobial and antitumoral immune defence mechanism (Sperner-Unterweger et al. 2010). Deprivation of thryptophan limits malignant tumour cells through an immune response mechanism, as well as inhibiting T-cell proliferation. Furthermore kynurenin, a catabolite of thryptophan, may contribute to T-cell apoptosis (Sperner-Unterweger et al 2010). A decrease in thryptophan concentration in conjunction with an increase in Kynurenin levels is considered as a biomarker of immune activation. This ratio is positively related to neopterin which is also considered as an inflammatory biomarker (Sperner-Unterweger et al 2010). Furthermore, vitamin B6 has been inversely linked to inflammatory biomarkers and oxidative stress (Shen et al 2010).

We applied structural equation models (SEM) to estimate three latent variables in the pathway to LC status, each one representing one of the mechanisms: DNA methylation, DNA synthesis and immune activation. It was assumed that specific biomarkers are related to each other and together they compose a latent variable (not measured directly) representing these mechanisms. In order to further evaluate how these three pathways might influence LC risk we have conducted a study using the dataset from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. We have developed a simplified model including plasma

and serum measurements of several key factors and evaluated their association with risk. As tobacco exposure is an important cause of LC, we considered it as an independent factor in LC causation. Age and cotinine were also included as independent factors for each latent variable because it is hypothesized that these three mechanisms may be affected by them.

MATERIALS AND METHODS

Study subjects

The EPIC study is a multi-centre study involving 10 European Countries and 23 centres (Riboli et al 2002). Amongst the 519,978 subjects, 2,206 were diagnosed with LC after an average of 10 years of follow-up. The LC cases were defined on the basis of the International Classification of Diseases for Oncology, Second Edition, and included all invasive cancers coded as C34. Cases with no available blood sample, date of blood collection or history of another cancer at the time of blood donation were excluded. For each case, two controls were randomly matched considering: country, gender, date of blood collection and date of birth (Johansson 2010). Cases with no controls were excluded from the sample. From the nested case-control, 39 subjects were excluded because of missing information from laboratory analyses or because their matched counterpart(s) had missing information. The final dataset, with complete information on folate, serine, vitamins B6 and B12, cotinine, homocysteine, methionine, neopterin, and kynurenin/tryptophan ratio (KTR), has 2638 subjects (891 cases and 1747 controls; 35 subjects with 1:1 matching and 856 with 1:2).

Biochemical analyses

Measurements of plasma cotinine and serum of serine, glycine, vitamin B6, folate, vitamin B12, vitamin B2, total homocysteine, methionine, neopterin, kynurenin and tryptophan were performed at Bevital A/S (www.bevital.no), blinded as to the case-control status (Johansson et al 2010).

Model assignment and statistical methods

A SEM with latent variables was applied to consider measurement and structural model simultaneously. Latent variable is defined as a random variable which cannot be estimated directly (measurement model) in which the basic premise is that the co-variation observed among the manifest variables is due to their relationship with the latent variable (McCutcheon 1987). Therefore, latent variable is estimated based on manifest variables which are observable (Skron dal and Rabe-Hesketh 2004). We considered three latent variables: DNA methylation (F1), DNA Synthesis (F2) and Immune activation (F3).

Based on the literature we considered the following observed variable to be represented as constructs for each latent:

1. *DNA Methylation* (F1): methionine (met), homocysteine (hcy), folate (fol), serine (ser) and Vitamin B12 (cobalamin) (cob);
2. *DNA Synthesis* (F2): Vitamin B6 (PLP), folate (fol), Vitamin B2 (riboflavin) (rib) and serine (ser) (or glycine (gly));
3. *Immune activation* (F3): Vitamin B6 (PLP), neopterin (npt) and kynurenin/tryptophan ratio (KTR);

Due to the importance of folate reactions on the DNA methylation branch, we checked the indirect effect of DNA synthesis (folate cycle, F2) on LC risk through DNA methylation (F1). We also included cotinine as a biomarker of tobacco exposure (active and passive short term exposure) as an independent cause of LC and as a factor for all latent variables. We propose the theoretical pathway shown in Figure 1: each square represents manifest variables whereas the circles represent latent ones. Each arrow represents one causal relation.

All analyses were conducted in Mplus 6. The matching design between cases and controls was handled by designating the stratum and setting complex sample for the final *probit* model. We used weighted least square with mean and variances adjusted (WLSMV) estimation (Muthén and Muthén, 2009). All observed variables were previously standardized (minus sample mean and divided by standard deviation) and considered with measurement errors which are represented by ϵ 's (Figure 1).

The following were used as fit indices: the comparative fit index (CFI; $>.90$ = considered as adequate), the Tucker-Lewis index (TLI; $>.90$ = considered as adequate) and root mean square error of approximation (RMSEA $<.08$ = good, $.05-.08$ = adequate, $.08-.10$ = marginal, $>.10$ = poor). The fit indices were evaluated in aggregate to draw conclusions about adequacy of fit.

RESULTS

In the current study, cases and controls presented the same mean age: 58 years. Amongst controls, 38.7% were never smokers and 37.0% were former smokers at the time of recruitment, compared to 10.8% and 28.8% respectively among the cases. The mean cotinine level was 231.07 nmol/L for controls and

896.43 nmol/L for cases. Amongst all participants 76.1% had normal serum homocysteine levels ($<15\mu\text{mol/L}$), 22.4% had mild, 1.4% moderate and 0.04% severe hyperhomocysteinemia (with corresponding ranges of 15-30 $\mu\text{mol/L}$, 30-100 $\mu\text{mol/L}$, and $>100\mu\text{mol/L}$, as defined by Seccareccia (2003)). In addition, 293 (11.1%) can be classified as folate deficient ($<7.5\text{nmol/L}$), and 199 (7.5%) as vitamin B12 deficient ($<200\text{pmol/L}$).

Table 1 shows four estimated model results: SEM-1 represents the hypothesized model in Figure 1, including the arrows with “X” and using Serine as a component of F1 and F2; SEM-2 represents a similar model dropping the direct effect of F2; SEM-3 represents the same hypothesized model in Figure 1 excluding hcy and ser from F1; SEM-4, Figure 1 considering arrows without “X”, is similar to SEM-3 replacing serine by glycine (in the folate cycle serine is converted to glycine).

Models SEM-1 and SEM-2 presented very high RMSEA and inadequate TLI. Model SEM-3 did not present a better fit, but replacing serine by glycine, model SEM-4 improved considerably. SEM-4 was selected as the final model with RMSEA=0.08, CFI=0.96 and TLI=0.93.

SEM-4 presented three latent variables. The F1 presented met and cob as positive factors whereas fol was negative and non significant. F2 presented four biomarkers positively related, in which PLP presented the highest weight. Finally, F3, as expected, presented PLP with positive weight and npt and KTR with negative weight, concluding that higher levels of F3 mean higher immune activity.

The effect of cotinine on LC risk was the highest, but it is interesting to note that, despite the fact that cotinine has significant effect on F2 and F3, its total indirect effect on LC is not significant ($p=0.388$). The impact of cotinine is restricted to the

levels of the latent variables F2 and F3, but the highest impact is observed directly to LC (the addition of one standard deviation in cotinine levels means an increase of 0.5 standard deviations in the probit score).

Interestingly the direct effect of F2 for LC was positive (0.26, $p=0.014$). Higher levels of F2 seem to be a direct risk to LC, but the indirect effect through F1 presents a higher and inverse effect (-0.31, $p=0.013$). In the end, the total effect of F2 is a non-significant reduction of risk (-0.05, $p=0.098$).

DISCUSSION

We have proposed a model, Figure 1, which includes three different mechanisms in LC risk in addition to tobacco exposure. We used information from subjects' blood samples as epidemiological biomarkers to represent these three intracellular metabolisms (DNA methylation-F1, nucleotide synthesis-F2 and immune activation-F3). We applied SEM with final outcome risk estimated by probit model.

Results in Table 1 show that the best fitted model is SEM-4 ($CFI>0.95$; $TLI>0.90$ and $RMSEA=0.08$). The three latent variables, F1, F2 and F3, where F1 did not include hcy, or ser, were in some way linked to LC risk (fol has no effect in this latent variable). Nucleotide synthesis, F2, did not have a final effect on LC, but presented significant (and with opposite signs) direct and indirect effects (the total effect lost significance). The indirect effect of F2 is about -0.31 SD, whereas the direct effect was 0.26 SD. This indirect effect means that the folate cycle has no total effect in LC, but this does have an effect in assisting DNA methylation. In F1 the weight of folate is small, non-significant and negative (-0.19) whereas the effect of

the folate cycle (F2) is positive and higher (0.47). The whole folate cycle, F2, has an important effect on methylation which in turn has a protective effect in LC. This effect from F2 means that it is positively related to SAM synthesis through methionine. The negative effect from folate in F1 we are interpreting as a negative relation with hcy levels. Even when hcy is dropped we still interpret F1 as the methionine-homocysteine re-methylation cycle. In some way folate may help to keep lower levels of hcy (Selhub 1999, 2006, Scott 1999).

The Met-Hcy re-methylation cycle is very complex; it involves other intracellular factors that we did not include. In this cycle SAM concentration is related to met and fol and depending on its concentration, MTHFR will or will not be activated. Its activation or inactivation decides if the intracellular folate will be used for methylation or kept non-methylated for DNA synthesis (Baluz et al 2002).

Serine is a very important one-carbon donor in the folate cycle as well as in the transsulfuration pathway (Gregory III et al. 2000), and its effect was identified in SEM-4 with glycine significant and positively related to F2. In Table 1 glycine presents the lowest weight in F3, but still important (0.23, $p < 0.001$).

The methylation pathway presented protection against LC with a reduction of 0.36 SD in the probit score. We do not have information on gene methylation: we are assuming that this latent F1, based on some co-factors needed for methylation, represents methylation. If F1 can be interpreted as methylation latent, this result is in agreement with previous studies in which methylation protects against cancer (Das and Signal 2004, Baluz et al 2002, Kim 2003).

Furthermore, we have found a reduction in LC risk via the immune activation pathway with a coefficient of -0.07 ($p = 0.037$). This pathway represents immune

activation and inflammatory markers because if npt or KTR are high, the score is low and if PLP is low, the score is high. In this way, high levels of F3 mean that PLP has adequate levels representing low inflammation and low oxidative stress (Shen et al 2010) and macrophage activation immune response is working well (Spencer et al 2010, Sperner-Unterweger et al 2010).

Finally, cotinine presented a positive and significant direct effect on LC, in which one extra SD in Cot means a 0.50 SD increase in LC likelihood. This result is in agreement with previous studies in which cotinine, passive tobacco exposure or smoking status presented significant and positive association with LC risk (McComack et al. 2010, Vineis et al. 2004, 2005, 2007, Taylor et al. 2007).

CONCLUSION

In conclusion, our results support a role of DNA methylation and immune activation in LC aetiology. Factors representing nucleotide synthesis also showed some association with LC risk, but these associations were weaker. Tobacco effect remains as the factor with the highest impact on lung cancer.

REFERENCES

- Ames BN. 2006. Low micronutrient intake may accelerate the degenerative diseases of aging through allocation of scarce micronutrients by triage. *PNAS*. 103 (47):17589-17594.
- Baluz K, do Carmo MGT, Rosas G. 2002. The role of folic acid on oncologic prevention and intervention: review. *Revista Brasileira de Cancerologia*, 2002, 48(4): 597-607.
- Brennan P, Hainaut P, Boffetta P. Genetics of lung-cancer susceptibility. 2010. *The Lancet Oncology*, Early Online Publication, 14 October 2010. doi:10.1016/S1470-2045(10)70126-1
- Das PM, Signal R. 2004. DNA methylation and cancer. *J Clin Oncol*. 22:4632-4642.
- Davis SR, Quinlivan EP, Stacpoole PW, Gregory III JF. 2006. Plasma glutathione and cystathionine concentrations are elevated but cysteine flux is unchanged by dietary vitamin B6 restriction in young men and women. *J Nutr*. 136:373-378.
- Eden A, Gaudet F, Waghmare A, Jaenish R. 2003. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* 300: 455.
- Feinberg AP, Vogelstein B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* 301:89-92
- Fenech M. 2010. Folate, DNA damage and the aging brain. *Mech Ageing Dev*. 131(4):236-41.
- Gregory III JF, Cuskelly GJ, Shane B et al.. 2000. Primed, constant infusion with [²H₃]serine allows in vivo kinetic measurement of serine turnover, homocysteine remethylation, and transsulfuration processes in human one-carbon metabolism. *Am J Clin Nutr*. 72:1535-1541.
- Hartman TJ, Woodson K, Stolzenberg-Solomon R, et al. 2001. Association of the b-vitamins pyridoxal 5'-phosphate (b6), B12, and folate with lung cancer risk in older men. *Am J Epidemiol.*; 153(7):688-694.
- Johansson M, Relton C, Ueland PM, et al. 2010. Serum B vitamin levels and risk of lung cancer. *JAMA*. 303(23):2377-2385.
- Kim Y-I. 2003. Role of folate in colon cancer development and progression. *J Nutr*. Nov;133(11 Suppl 1):3731S-3739S.
- Lengauer C. 2003. An Unstable Liaison. *SCIENCE*. Vol 300: 442-443.
- McCormack V, Agudo A, Dahm CC, Overvad K, Olsen A5, Tjønneland A, et al. 2010. Cigar and pipe smoking and cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Int. J. Cancer*: 127, 2402-2411.
- McCutcheon AL. 1987. Latent class analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-064. Newbury Park, CA.
- Muthén B, Muthén L. Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén; 1998-2010.
- Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002; 5:1113-1124.
- Scott JM. 1999. Folate and vitamin B12. *Proc Nutr Soc*. 58(2):441-448.

- Seccareccia F, Zuccaro P, Pacifici R, et al. Serum cotinine as a marker of environmental *tobacco* smoke exposure in epidemiological studies: the experience of the MATISS Project. *Eur J Epidemiol*. 2003; 18(6):487-492.
- Selhub J. 1999. Homocysteine metabolism. *Annu. Rev Nutr.*; 19:217-246.
- Selhub J. 2006. The Many Facets of Hyperhomocysteinemia: Studies from the Framingham Cohorts. *J. Nutr.* 136 (6):1726S-1730S
- Shen J, Lai C-Q, Mattei J, Ordovas JM, Lucker KL. 2010. Association of vitamin B-6 with inflammation, oxidative stress, and chronic inflammatory conditions: the Boston Puerto Rican Health Study. *Am J Clin Nutr.* 91:337-342.
- Skrondal A, Rabe-Hesketh S. 2004. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Chapman & Hall/CRC interdisciplinary statistics series. Washington, DC.
- Spencer ME, Jain A, Matteini A, Beamer BA, Wang N-Y, Leng SX et al. 2010. Serum Levels of the Immune Activation Marker Neopterin Change With Age and Gender and Are Modified by Race, BMI, and Percentage of Body Fat. *J Gerontol A Biol Sci Med Sci.* 65(8):858–865.
- Sperner-Unterwehgera B, Neurauter G, Klieber M, Kurzc K, Meranera V, Zeimet A, Fuchs D. 2010. Enhanced tryptophan degradation in patients with ovarian carcinoma correlates with several serum soluble immune activation markers. *Immunobiology* xxx (2010) xxx–xxx. doi:10.1016/j.imbio.2010.07.010
- Stover PJ. One-carbon metabolism-genome interactions in folate-associated pathologies. *J Nutr.* 2009. 2402-2405.
- Taylor R, Najafi F, Dobson A. 2007. Meta-analysis of studies of passive smoking and lung cancer: effects of study type and continent. *Int J of Epidemiol* 36:1048–1059.
- Vineis P, Alavanja M, Buffler P, Fontham E, Franceschi S, Gao YT, et al. 2004 Tobacco and cancer: recent epidemiological evidence. *J Natl Cancer Inst* 96(2):99-106.
- Vineis P, Airoldi L, Veglia P, Olgiati L, Pastorelli R, Autrup H, et al. 2005. Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ* 330(7486):277-281.
- Vineis P, Hoek G, Krzyzanowski M, Vigna-Taglianti F, Veglia F, Airoldi L, et al. 2007. Lung cancers attributable to environmental tobacco smoke and air pollution in non-smokers in different European countries: a prospective study. *Environmental Health* 6:7.
- Visnes T, Doseth B, Pettersen HS, Hagen L, Sousa MML, Akbari M, et al. 2009. Uracil in DNA and its processing by different DNA glycosylases. *Phil. Trans. R. Soc. B* 2009 364, 563-568 doi: 10.1098/rstb.2008.0186
- Ulrich CM, Reed MC, Hight HF. 2008 Modeling folate, one-carbon metabolism, and DNA methylation. *Nutrition Reviews*.66(1): S27-S30.

Figure 1. Theoretical pathway considering cotinine, and three latent variables as causes of lung cancer with the latent DNA synthesis as direct and indirect cause.

Table 1. SEMs results for lung cancer probit model

	SEM-1		SEM-2		SEM-3		SEM-4	
	Standardized estimate	p-value	Standardized estimate	p-value	Standardized estimate	p-value	Standardized estimate	p-value
LC								
F1	0.03	0.476	0.03	0.510	-0.36	0.022	-0.36	0.018
F2	-0.04	0.198			0.27	0.047	0.26	0.041
F3	-0.09	0.006	-0.09	0.004	-0.07	0.040	-0.07	0.037
Cot	0.52	<0.001	0.54	<0.001	0.51	<0.001	0.50	<0.001
F1								
AG	-0.21	<0.001	-0.21	<0.001	-0.23	0.001	-0.23	<0.001
COT	-0.20	<0.001	-0.20	<0.001	-0.02	0.802	-0.04	0.546
F2	0.04	0.161	0.04	0.166	0.89	<0.001	0.86	<0.001
F2								
AG	0.08	0.008	0.09	0.005	0.07	0.024	0.08	0.005
COT	-0.28	<0.001	-0.28	<0.001	-0.28	0.000	-0.27	<0.001
F3								
AG	-0.33	<0.001	-0.33	<0.001	-0.32	<0.001	-0.32	<0.001
COT	0.09	0.003	0.09	0.003	0.10	0.004	0.09	0.004
Latent -F1								
Hcy	-0.65	-	-0.65	-				
Met	-0.15	<0.001	-0.15	0.000	0.23	-	0.25	-
Fol	0.19	<0.001	0.19	0.000	-0.22	0.063	-0.19	0.076
Cob	0.22	<0.001	0.22	0.000	0.16	0.000	0.17	<0.001
Ser	-0.33	<0.001	-0.33	0.000				
Latent -F2								
Fol	0.29	-	0.29	-	0.49	-	0.47	-
PLP	0.63	<0.001	0.63	<0.001	0.63	<0.001	0.64	<0.001
Rib	0.43	<0.001	0.43	<0.001	0.43	<0.001	0.42	<0.001
Ser	0.16	<0.001	0.16	<0.001	0.09	<0.001		
Gly							0.23	<0.001
Latent -F3								
PLP	0.14	-	0.14	-	0.13	-	0.13	-
Npt	-0.65	<0.001	-0.65	0.000	-0.65	<0.001	-0.65	<0.001
KTR	-0.66	<0.001	-0.66	0.000	-0.66	<0.001	-0.66	<0.001
Effects from F2 to LC								
Total	-0.04	0.214	0.001	0.547	-0.06	0.092	-0.05	0.098
Total indirect	0.001	0.523	0.001	0.547	-0.32	0.015	-0.31	0.013
Effects from COT to LC								
Total	0.521	<0.001	0.52	<0.001	0.52	<0.001	0.52	<0.001
Total indirect	-0.004	0.781	-0.01	0.117	0.02	0.537	0.02	0.388

	SEM-1		SEM-2		SEM-3		SEM-4	
	Standardized estimate	p- value	Standardized estimate	p- value	Standardized estimate	p- value	Standardized estimate	p- value
Specific indirect								
LC<F1<-F2	0.001	0.523	0.001	0.547	-0.32	0.015	-0.31	0.013
LC<F1<COT	-0.006	0.476	-0.005	0.510	0.01	0.801	0.02	0.549
LC<F2<COT	0.011	0.206			-0.08	0.075	-0.07	0.064
LC<F3<COT	-0.009	0.017	-0.009	0.014	-0.01	0.061	-0.01	0.060
LC<F1<F2<COT	0.000	0.522	0.000	0.546	0.09	0.030	0.08	0.026
Model Fit								
CFI	0.91		0.91		0.92		0.96	
TLI	0.86		0.87		0.86		0.93	
RMSEA	0.13		0.12		0.11		0.08	

5. CONSIDERAÇÕES FINAIS

Do manuscrito 1 concluí-se que há evidências de que a cotinina é um bom biomarcador de exposição ao tabaco no curto prazo. É claro que os níveis de cotinina em um ex-fumante estarão baixos se comparados a um fumante, o que justifica a utilização de questões sobre o histórico de tabagismo. Desse manuscrito, foi possível identificar a importância da cotinina para avaliação dos efeitos da exposição ao fumo passivo: indivíduos não fumantes apresentaram níveis consideravelmente elevados de cotinina no sangue quando expostos passivamente ao tabaco. Essa diferença não foi significativa na comparação entre nunca e ex-fumantes, mas o valor médio da diferença foi relativamente elevado (3,61 nmol/L). Ressaltamos a importância de que novos estudos investiguem a hipótese de que ex-fumantes possam estar mais suscetíveis à exposição passiva, devido a alterações genéticas que possam ter ocorrido na época em que eram fumantes.

Os manuscritos 2 e 3 abordam de maneira diferente as interrelações de componentes do ciclo do carbono e da ativação da imunidade (e transulfuração no manuscrito 2) no risco de câncer de pulmão. Nos dois manuscritos, a cotinina se mostrou como uma variável de exposição à nicotina com efeito bastante elevado e com significativo aumento do risco de câncer. A hipótese levantada sobre os efeitos da metilação como proteção contra o câncer (via metionina, ou via a variável latente de metilação) foi confirmada nas duas abordagens.

No manuscrito 2, onde foram utilizadas duas escalas diferentes para as variáveis de exposição, ficaram dúvidas com relação à importância da síntese de nucleotídeos e da via de ativação da imunidade na causação do câncer. Por outro lado, no

manuscrito 3, a hipótese de que uma maior exposição à inflamação (ou menor ativação da imunidade) sugere um maior risco de câncer, foi estatisticamente confirmada. Com relação à síntese de nucleotídeos, no manuscrito 3, não confirmou-se a hipótese de relação com o câncer de pulmão, mas confirmou-se a hipótese de que o ciclo do folato (F2) é importante preditor do ciclo da homocisteína-metionina (F1, metilação).

Além das conclusões dos manuscritos é importante ressaltar que a metodologia de MEE empregada, permitiu confirmar a presença cofatores causais que um modelo de regressão tradicional talvez não identificasse. Por exemplo, no MEE do manuscrito 2, com base na análise sem a transformação logaritmica, é possível verificar que a vitamina B6 não apresenta uma relação direta com o câncer de pulmão, mas é um importante cofator na metilação do DNA, apresentando assim, um efeito indireto na redução do risco do câncer. Os cofatores enzimáticos e os transportadores de carbono nem sempre aparecem como fatores significativos em regressões tradicionais, apesar de apresentarem uma forte base teórica bioquímica dentro dos ciclos metabólicos. Além disso, é importante que se verifique, dentro dos ciclos metabólicos, onde esses cofatores apresentam maior impacto; em uma regressão tradicional essa posição não poderia ser considerada.

Por fim, ressalta-se que a abordagem de MEE com variáveis latentes permite uma visão mais abrangente, sem considerar as posições exatas de cada causa dentro dos ciclos metabólicos, mas como fatores que estão associados, e permitindo que vias metabólicas possam ter seus efeitos estimados. Nessa abordagem, foi possível identificar que as variáveis envolvidas na síntese de nucleotídeos, conjuntamente, não apresentam um efeito na redução do risco de câncer de pulmão, mas apresentam uma

função importante para o estímulo da via com os componentes que permitirão a metilação, que por sua vez, ajudam a prevenir o câncer.

Outros fatores podem estar envolvidos nesses ciclos, como é o caso das enzimas envolvidas nos ciclos do folato e da metilação. Um estudo posterior poderia verificar se, por exemplo, algum polimorfismo genético influencia esses ciclos, e se mudariam os resultados que foram encontrados nesta tese.

6. REFERÊNCIAS

- Alencar GP. Influência dos fatores de situação sócio-econômica, de aceitação da gravidez e da assistência pré-natal na mortalidade fetal: análise com modelagem de equações estruturais [tese de doutorado]. São Paulo: Faculdade de Saúde Pública da USP; 2009.
- Assmann SF, Hosmer DW, Lemeshow S, et al. Confidence intervals for measures of interaction. *Epidemiology*. 1996;7(3):286–290.
- Baluz K. Do Carmo MGT, Rosas G. O papel do ácido fólico na prevenção e na terapêutica oncológica: revisão. *Rev Bras de Cancerol* 2002; 48(4):597-607.
- Blot JW. Synergism and interaction: are they equivalent? [carta]. *Am J Epidemiol*. 1979;110(1):99-100.
- Bollen KA, Long JS. Testing Structural equation models. A SAGE Focus Ed. 1993.
- Das PM, Signal R. 2004. DNA methylation and cancer. *J Clin Oncol*. 22:4632-4642.
- Davis SR, Quinlivan EP, Stacpoole PW, Gregory III JF. Plasma glutathione and cystathionine concentrations are elevated but cysteine flux is unchanged by dietary vitamin B6 restriction in young men and women. *J Nutr*. 2006. 136:373-378.
- Dragow F. Polychoric and polyserial correlation. In: Kotz L e Johnson NL, coordenadores. *Encyclopedia of statistical Sciences*. Nova Iorque: John Wiley; 1988. Vol. 7, p. 69-74.
- Fenech, M. Folate, DNA damage and aging brain. *Mech Ageing Dev*. 2010. 131(4):236-241.

- Ferraz CM, Steluti J, Marchioni DML. As vitaminas e minerais relacionados à estabilidade genômica e à proteção ao câncer. *Nutrire: rev. Soc. Bras. Alim. Nutr.* = J. Brazilian Soc. Food Nutr 2010.35(2):181-199.
- Greenland S. Interaction in epidemiology: relevance, identification, and estimation. *Epidemiology* 2009. 20(1): 14-17.
- Greenland S, Brumback B. An overview of relations among causal modeling methods. *Int J Epidemiol* 2002; 31:1030-1037.
- Greenland S, Peal J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10:37-48.
- Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. *Scand J Work Environ health*. 1988. 14: 125-129.
- Harris B. Tetrachoric correlation coefficient. In L. Kotz e NL Johnson, coordenadores. *Encyclopedia of statistical Sciences*. 1 ed. Nova Iorque: John Wiley;1988. Vol. 9, p. 223-225.
- Hoffmann K, Heidemann C, Weikert C, Schulze MB, Boeing H. Estimating of disease due to classes of sufficient causes. *Am J Epidemiol* 2005;163(1):76-83.
- Hogan MD, Kupper LL RE. Estimation versus detection in the assessment of synergy. *Am J Epidemiol*. 1978;108(2):159-60.
- Hogan MD, Kupper LL, Most BM, Haseman JK. Alternatives to Rothman's approach for assessing synergism (or antagonism) in cohort studies. *Am J Epidemiol*. 1978;108(1):60-67.
- Hosmer, D.W., & Lemeshow, S. *Applied Logistic Regression*. 1989; New York: Wiley.

- Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. *Epidemiology*. 1992;3(5):452–456.
- IARC-International Agency for Research on Cancer. GLOBOCAN 2008- Cancer Fact Sheet. [acessao em 03/12/2010]. Disponível em: <http://globocan.iarc.fr/factsheets/cancers/lung.asp>
- Johansson M, Relton C, Ueland PM, et al. Serum B vitamin levels and risk of lung cancer. *JAMA*. 2010; 303(23):2377-2385.
- James LR, Mulaik SA, Brett JM. Causal analysis, assumptions, models, and data. 1982. Beverly Hills, CA Sage.
- Ji Y-N, Zhan P, Wang J, Qiu L-X, Yu L-K. APE1 Asp148Glu gene polymorphism and lung cancer risk: a meta-analysis. *Mol Biol Rep* 2010. Doi: 10.1007/s11033-010-0584-2.
- Jöreskog KG. Structural equation modeling with ordinal variables using LISREL [monografia na internet] Scientific Software International. Chicago, 2002. [acesso em 02 dez 2007]. Disponível em: <http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf>
- Kenny DA. Correlation and causality. 1979. New York: Wiley.
- Kraemer HC. Biserial correlation. In: L. Kotz e NL Johnson, coordenadores. *Encyclopedia of statistical Sciences*. Nova Iorque: John Wiley; 1982. Vol. 1,p. 276-280.
- Kupper LL, Hogan MD. Interaction in epidemiologic studies. *Am J Epidemiol*. 1978;108(6):447-453.

- Little RJA, Rubin DB. Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches. *Annual Review of Public Health* 2000; 21: 121-145.
- Lo Y-L, Hsiao C-F, Jou Y-S, Chang G-C, Tsai Y-H, Su W-C, et al. Polymorphisms of MLH1 and MSH2 genes and the risk of lung cancer among never smokers. *Lung Cancer*. 2010. Doi:10.1016/j.lungcan.2010.10.009.
- Long JS. Confirmatory factor analysis. Series: Quantitative applications in the Social Sciences. SAGE Univ Press. 1983.
- MacMahon B, Pugh TF. Causes and entities of disease. In: Clark DW, MacMahon B (eds). *Preventive Medicine*. Boston, 1967.
- Maldonado G, Greenland S. Estimating Causal Effects. *Int J Epidemiol* 2002; 31: 422-429.
- Melhado TT. Influência local para modelos de equações estruturais com distribuição elíptica [tese de doutorado]. São Paulo: Instituto de Matemática e Estatística da USP; 2009.
- Mackie JL. Causes and conditions. *Am Phil Quar*. 1965; 2 (4): 245-64.
- Mackie JL. *The cement of the universe: A study of causation*. Oxford: Oxford University Press, 1974.
- Marchioni DML, de Lima FEL, Fisberg RM. 2008. Dietary patterns and risk of breast cancer: a case-control study in the Northeast of Brazil. *Nutrire: rev.Soc. Bras. Alim. Nutr.=J. Brazilian Soc. Food*: 33 (1):31-42.
- Miettinen O. Confounding and effect-modification. *Am J Epidemiol*. 1974;100(5):350-53.

- Miettinen O. Causal and preventive interdependence. Elementary principles. *Scand J Work Environ Health*. 1982. 8(3):159-68.
- Muthén B, Satorra A. Technical aspects of muthén's liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika* 1995. 60(4):489-503.
- Niles HE. Correlation, causation and Wright's theory of path coefficients. *Genetics* 1922; 7: 258-273
- Pearl J. Causal diagrams for empirical research. *Biometrika* 1995. 82(4): 669-710.
- Pearl J. *Causality: models, reasoning, and inference*. New York: Cambridge University Press 2000.
- Poole C. Positivized epidemiology and the model of sufficient and component causes. *Int J Epidemiol* 2001; 30:707-709.
- Retherford RD, Choe MK. *Statistical models for causal analysis*. New York: Wiley. 1993.
- Richardson DB, Kaufman J. Estimation of the relative excess risk due to interaction and associated confidence bounds. *Am J Epidemiol* 2009; 169:756-760.
- Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002; 5:1113-1124.
- RL-UFL - Robertson Laboratory at the University of Florida. Research Interests. [acesso em 30 jan 2011] Disponível em: <http://www.med.ufl.edu/biochem/keithr/>
- Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001; 12(3):313-20.

- Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11(5):550-60.
- Rosenbaum PR. Optimal Matching for Observational Studies. *J Am Stat Ass* 1989;84(408):1024–32.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; 39(1):33-38.
- Rothman KJ. Synergism and antagonism in cause-effect relationships. *Am J Epidemiol.* 1974;99(6):385-88.
- Rothman KJ. The estimation of synergism or antagonism. *Am J Epidemiol.* 1976a;103(5):506-11.
- Rothman KJ. Causes. *Am J Epidemiol.* 1976b;104:587-92.
- Rothman KJ. *Modern Epidemiology*. Boston: Little Brown, 1986.
- Rothman KJ. *Epidemiology an introduction*. Oxford Univ Press, 2002.
- Skron dal A. Interactions as departure from additivity in case-control studies: a cautionary note. *Am J Epidemiol* 2003; 158(3):251-258.
- Skron dal A, Rabe-Hesketh S. *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Florida: CRC. 2004.
- Stefani ED, Boffeta P, Ronco AL, Deneo-Pellegrini H, Acosta G, Gutiérrez LP et al. Nutrient patterns and risk of lung cancer: a factor analysis in Uruguayan men. *Lung Cancer*. 2008. 61:283-291.
- Stover PJ. One-carbon metabolism-genome interactions in folate-associated pathologies. *J Nutr.* 2009. 2402-2405.

- Ulrich CM, Reed MC, Hijhout HF. Modeling folate, one-carbon metabolism, and DNA methylation. *Nutrition Reviews*. 2008; 66(1): S27-S30.
- Visnes T, Doseth B, Pettersen HS, Hagen L, Sousa MML, Akbari M, et al. Uracil in DNA and its processing by different DNA glycosylases. *Phil. Trans. R. Soc. B* 2009 364, 563-568 doi: 10.1098/rstb.2008.0186
- VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* 2009; 20(1): 18-26.
- VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-cause framework. *Epidemiology* 2007a; 18(3): 329-339.
- VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol* 2007b; 166:1096-1104.
- VanderWeele TJ, Robins JM. Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika* 2008; 95(1): 49-61.
- VanderWeele TJ, Vansteelandt S, Robins JM. Marginal Structural models for sufficient cause interactions. *Am J Epidemiol* 2010; 171(4): 506-514.
- WHO - World Health Organization. The smokers body -2006. Geneva; 2006 [acesso em 1 dez 2010] Disponível em: http://www.who.int/entity/tobacco/resources/publications/smokersbody_en_fr.pdf
- WCRF/AICR - World Cancer Research Fund/American Institute for Cancer Research. Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective. Washington DC: AICR, 2007.
- Wright S. Correlation and Causation. *J Agric Res* 1921; 20:557-585.

YU CH. Causation in Quantitative Research Methodologies from Path Modeling, SEM to TETRAD. Theory & Science. 2007 [acesso em 03/12/2010].
Disponível em: <http://theoryandscience.icaap.org/content/vol9.3/chong.html>

ANEXO A

Medidas de Sinergia

Este anexo apresenta uma síntese de várias medidas de sinergia propostas por diferentes autores desde 1976, quando Rothman apresentou pela primeira vez o seu modelo causal.

Segundo ROTHMAN (1976a), a sinergia é a ocorrência de um excesso de efeito de risco, devido à interação (interação “biológica” positiva); quando há uma diminuição desse risco, o autor a denomina *antagonismo* (interação “biológica” negativa), similarmente ao que MIETTINEN(1974) chamou de *efeito modificador*. BLOT (1979) ressaltou que essa idéia de sinergia é equivalente a um tipo específico de interação, aquela que considera um excesso de risco relativo, para o caso de um modelo aditivo, ou seja, a exposição a dois ou mais fatores resulta em um número maior de casos de doença, sendo esse número maior do que a simples soma dos fatores separadamente. No entanto, se o modelo adotado for multiplicativo (por exemplo, a regressão logística), essa sinergia que ocorreria no modelo aditivo pode não ser identificada. Então, a percepção da sinergia depende do modelo estatístico empregado.

O modelo causal de Rothman com três conjuntos suficientes está representado na Figura A1. A causa componente A é uma causa necessária, pois está presente nas três alternativas possíveis de conjuntos suficientes. Nesse modelo, é importante ressaltar que uma doença pode ser determinada por mais de um mecanismo causal, e cada mecanismo envolve a ação de múltiplas causas componentes. Além disso, as causas componentes não precisam acontecer num mesmo momento no tempo.

Conjunto Suficiente I Conjunto Suficiente II Conjunto Suficiente III

Figura A1 - Esquematização conceitual de três conjuntos suficientes para uma doença (ROTHMAN, 1976b)

Em 1974, MIETTINEN discutiu os conceitos de confundimento e de efeito modificador. Nesse mesmo ano, ROTHMAN (1974) introduziu a sua ideia sobre sinergia e antagonismo. Um fator pode apresentar correlação à exposição e também ao efeito (à doença estudada), e isso pode significar confundimento. MIETTINEN (1974) ressalta sua preocupação com esses dois conceitos nos modelos usados em epidemiologia. Sua sugestão é que fatores de confundimento sejam incluídos não como causas, mas como controle para esse confundimento (indicador de risco não-causal). Por outro lado, se há uma associação, controlando-se confundimentos, entre um fator com efeitos que diferem entre estratificações de outro fator, ou seja, se não há uma homogeneidade de riscos, é necessário estudar os efeitos de interação.

Considerando o modelo de ROTHMAN (1976a, 1976b, 1986) para apenas duas causas componentes A e B, e juntando as demais causas, supostamente desconhecidas em U, tem-se o esquema da Figura A2. Esse esquema representa quatro classes de conjuntos suficientes: os que têm a interação entre A e B, os que têm somente A (agindo de forma independente de B), os que têm somente B (agindo

de forma independente de A) e os que não têm A nem B (possuindo apenas causas desconhecidas). As pessoas que desenvolveram a doença e estiveram expostas às causas componentes A e B podem ter desenvolvido a doença por qualquer um desses quatro conjuntos suficientes da Figura A2. Para algumas pessoas, a doença pode ter sido causada pela ação integrada das causas componentes A e B, mas outras que estiveram expostas a essas causas podem ter desenvolvido a doença por outros mecanismos causais, alguma das outras três representações da Figura A2 (quais sejam por ação isolada de A ou de B, ou mesmo sem nenhuma ação dessas causas componentes, devido exclusivamente às causas desconhecidas).

Figura A2 - Esquematização conceitual de quatro classes envolvendo duas causas componentes para uma doença (ROTHMAN, 1986)

As informações disponíveis em estudos epidemiológicos referem-se à exposição das pessoas que adoeceram (ou não) com relação aos diferentes fatores de risco. Essas exposições não necessariamente causaram a doença, pois algumas pessoas não são suscetíveis a alguns desses fatores.

Desse modo, o risco de doença devido à interação “biológica” entre as duas causas componentes A e B pode ser estimado pela expressão:

$$I(AB) = [R_{AB} - R_U] - [R_A - R_U] - [R_B - R_U] = R_{AB} - R_A - R_B + R_U, \quad (\text{A.1})$$

onde R_{AB} é o risco observado para os indivíduos sujeitos às duas exposições, A e B; R_A é o risco observado naqueles com exposição somente ao fator A; R_B é o risco observado somente entre os que tiveram exposição à B; e, finalmente, R_U é o risco observado naqueles que não se expuseram ao fator A nem ao B.

Esse risco $I(AB)$ é atribuído à interação das causas A e B. Se o valor dessa expressão for zero, pode-se dizer que as causas A e B são “biologicamente” independentes. Caso contrário, há sinergia entre elas.

Em situação de independência biológica das causas componentes A e B, tem-se que $I(AB) = 0$. Então, observa-se a condição de aditividade dos efeitos absolutos:

$$[R_{AB} - R_U] = [R_A - R_U] + [R_B - R_U]. \quad (\text{A.2})$$

Isso significa que o efeito conjunto dos dois fatores é igual à soma dos efeitos de cada um separadamente, ou seja, existe independência estatística dos dois fatores (para um modelo aditivo).

Essa forma de medir a interação entre as causas componentes foi bastante trabalhada por diferentes pesquisadores. ROTHMAN (1976a) propôs uma medida dessa sinergia. Essa medida é a simples razão entre o efeito da interação observada dividido pelo efeito esperado sob hipótese de independência. O efeito da interação observada é o risco observado para expostos aos dois fatores menos o risco observado naqueles sem ambos os fatores (ROTHMAN, 1976a). O risco esperado sob hipótese de independência é obtido por $R_A - R_A R_C + R_B - R_B R_C - R_A R_B + R_A R_B R_C$ da teoria de probabilidade.

Então, considerando os mesmos dois fatores de risco A e B, temos $R_{11}=R_{AB}$, $R_{00} = R_U$, $R_{01} = R_A + R_U - R_A R_U$ e $R_{10} = R_B + R_U - R_B R_U$, onde os valores de R_{11} , R_{00} , R_{01} e R_{10} são obtidos nos estudos epidemiológicos. Assim, a sinergia é dada por:

$$S = \frac{R_{11} - R_{00}}{(R_{10} - R_{00}) + (R_{01} - R_{00}) - \frac{(R_{10} - R_{00})(R_{01} - R_{00})}{(1 - R_{00})}}, \quad (\text{A.3})$$

Quando os efeitos são pequenos, ROTHMAN (1976a) simplifica essa fórmula da seguinte maneira:

$$S' = \frac{R_{11} - R_{00}}{(R_{10} - R_{00}) + (R_{01} - R_{00})} = \frac{RR_{11} - 1}{RR_{10} + RR_{01} - 2} \quad (\text{A.4})$$

onde $RR_{ij} = R_{ij}/R_{00}$.

Essa estimativa de sinergia é obtida para fatores binários (presença ou ausência do fator de risco), mas, nesse mesmo artigo, ele também introduz uma estimativa para fatores com várias categorias (dados ordinais). Baseado numa série de Taylor de 1ª ordem, Rothman conseguiu estimar o erro padrão do logaritmo natural da estimativa de S. Considerando, para amostras grandes, a distribuição log-normal, chegou a boas estimativas de intervalo de confiança para estudos de coorte e de caso-controle (ROTHMAN, 1976a).

Em 1978, HOGAN et al. (HOGAN e KUPPER 1978, HOGAN et al. 1978) propuseram uma nova medida de sinergia para estudos de coorte considerando também causas binárias. No mesmo ano, KUPPER e HOGAN (1978) incluíram mais categorias para as causas. HOGAN et al. (1978) ressaltam, assim como Rothman, que há uma dependência potencial do valor da sinergia com a escala de medida que é empregada. Eles sugerem que a comparação feita entre o risco da interação observada dividida pelo efeito esperado sob hipótese de independência simplificada por Rothman para S' (equação A.4) pode ser feita montando a hipótese de outra maneira. Sob independência, temos aproximadamente:

$$(R_{11} - R_{00}) = (R_{10} - R_{00}) + (R_{01} - R_{00}) \quad \rightarrow \quad R_{11} - R_{10} - R_{01} + R_{00} = 0, \quad (\text{A.5})$$

onde R_{11} , R_{10} , R_{01} e R_{00} são os mesmos definidos anteriormente.

Então, pode-se considerar a estatística $T' = R_{11} - R_{10} - R_{01} + R_{00}$, que é a usual de análise de variâncias para avaliação do contraste de interação entre dois fatores (HOGAN et al, 1978), sendo que cada um dos fatores tem dois níveis apenas. Nota-se que essa estatística é igual à equação A.4, já trabalhada por Rothman. Então, testar $H_0: S'=1$ é equivalente a testar $H_0: T'=0$. Usualmente, para testar essa hipótese utiliza-se a estatística

$$Z' = \frac{\hat{T}' - 0}{\sqrt{\hat{V}(\hat{T}')}}, \quad (\text{A.6})$$

onde $\hat{T}' = \hat{R}_{11} - \hat{R}_{10} - \hat{R}_{01} + \hat{R}_{00}$ e $\hat{V}(\hat{T}') = \sum_{i=0}^1 \sum_{j=0}^1 \frac{\hat{R}_{ij}(1 - \hat{R}_{ij})}{N_{ij}}$, e N_{ij} é o número de

indivíduos na célula correspondente à exposição i e j . Assume-se que o número de indivíduos doentes entre os expostos é uma variável aleatória com distribuição $Binomial(N_{ij}, R_{ij})$ e que, sob a hipótese de H_0 ser verdadeira, a estatística Z' tem distribuição normal padrão para um tamanho razoavelmente grande de N_{ij} . A vantagem do uso de S' deve-se principalmente à familiaridade com uso de estatísticas de risco relativo, usuais em epidemiologia; mas, por outro lado, HOGAN et al. (1978) destacam não só que, em várias situações, S' não pode ser estimado (indefinição do logaritmo) mas também que os valores de p podem ser falsamente significativos, ou muito pequenos. A estatística T' , que foi criada para estudos de coorte também pode ser adaptada para estudos de caso-controle usando, no lugar de R , o RR (risco relativo, dividindo-se R_{ij} por R_{00}). Com base na distribuição $Binomial(N_{ij}, R_{ij})$, HOGAN et al. (1978) propuseram um teste de razão de verossimilhança, que tem maior poder estatístico, porém com uma dificuldade de

estimação, que depende de um função de maximização iterativa. Como alternativa, desenvolveram um teste que foi comparado com o da razão de verossimilhança, apresentando bons resultados, sendo estes melhores do que o teste em T'.

O teste que HOGAN et al. (1978) destacam como mais poderoso considera o risco para exposição simultânea sob hipótese de independência, R_{11}^0 (o evento U é considerado independente de A e de B), dada por:

$$R_{00} + (R_{10} - R_{00}) + (R_{01} - R_{00}) - \frac{(R_{10} - R_{00})(R_{01} - R_{00})}{(1 - R_{00})}, \text{ que pode ser reescrita da}$$

seguinte maneira: $R_{11}^0 = 1 - \frac{(1 - R_{10})(1 - R_{01})}{(1 - R_{00})}$. Então, mantém-se uma medida similar

a T', $T = R_{11} - R_{11}^0$. A estimativa de máxima verossimilhança para R_{11}^0 é

$$\hat{R}_{11}^0 = 1 - \frac{(1 - \hat{R}_{10})(1 - \hat{R}_{01})}{(1 - \hat{R}_{00})}, \text{ onde } \hat{R}_{ij} \text{ são os riscos, ou proporções, observados na}$$

amostra estudada. O teste para $H_0: T=0$ pode ser baseado na estatística $Z = \frac{\hat{T} - 0}{\sqrt{\hat{V}(\hat{T})}}$,

onde

$$\hat{T} = \hat{R}_{11} - \hat{R}_{11}^0 \quad \text{e} \quad (\text{A.7})$$

$$\hat{V}(\hat{T}) \equiv \frac{\hat{R}_{11}(1 - \hat{R}_{11})}{N_{11}} + \frac{(1 - \hat{R}_{10})(1 - \hat{R}_{01})}{(1 - \hat{R}_{00})^2} \left\{ \frac{\hat{R}_{00}(1 - \hat{R}_{10})(1 - \hat{R}_{01})}{(1 - \hat{R}_{00})N_{00}} + \frac{\hat{R}_{10}(1 - \hat{R}_{01})}{N_{10}} + \frac{\hat{R}_{01}(1 - \hat{R}_{10})}{N_{01}} \right\}. \quad (\text{A.8})$$

HOFFMANN et al. (2005) apresentam uma forma de quantificar os conjuntos suficientes do modelo teórico do Rothman com base nos resultados de uma regressão logística. Nessa regressão, optaram por incluir as interações estatísticas das causas binárias em até quatro graus, ou seja, o modelo que eles apresentaram tem quatro variáveis explicativas que podem não interagir com as demais, ou podem interagir

com mais uma, duas ou três variáveis (interação estatística para um modelo multiplicativo). O modelo resultante apresenta interações significativas em quatro situações: três delas com três variáveis e uma delas com quatro. Considerou-se, então, a identificação de quatro conjuntos suficientes (já que os demais fatores, interagindo ou não, não eram significativos) e a força da relação atribuída a cada um deles foi o risco relativo. Sendo assim, o conjunto suficiente com quatro variáveis foi o que apresentou maior força, apesar de não ser necessariamente o que tenha apresentado maior prevalência (os autores incluíram o cálculo da prevalência para cada conjunto suficiente). Nenhuma causa necessária pôde ser identificada, pois não houve uma causa componente que estivesse presente nos quatro conjuntos suficientes encontrados. Essa sugestão da análise das interações através das interações no modelo log-linear é bastante usual entre os estatísticos, porém há uma necessidade de se ter uma amostra bastante grande para que elas possam aparecer como significativas. Quando o evento é raro, ou quando a amostra não é suficientemente grande, as interações nesse tipo de modelo podem até confundir a análise dos resultados, e muitas vezes não são significativas, apesar de poderem existir.

Vários artigos foram publicados aplicando a ideia de modelos contrafactuais para o modelo de conjuntos suficientes de causas componentes (VANDERWEELE e ROBINS, 2007a 2007b, 2008, GREENLAND 2009, MIETTINEN 1982). MIETTINEN (1982) apresenta a formulação das respostas potenciais para o caso de duas variáveis binárias, indicando os vários tipos de possíveis respostas. Nesse caso, as duas variáveis de exposição, X_1 e X_2 , podem ser presentes ou ausentes, e a doença D representa a presença, ou não, da doença. Então, no modelo contrafactual, as possíveis combinações de exposição são: (1,1), (0,1), (1,0) e (0,0). Para cada

indivíduo, segundo MIETTINEN (1982), 16 (2^4) tipos de respostas potenciais são possíveis. Por exemplo, se $D_{11}(w)=D_{01}(w)=D_{00}(w)=0$ e $D_{10}(w)=1$, ou se $D_{11}(w)=D_{10}(w)=D_{00}(w)=0$ e $D_{01}(w)=1$, ou se $D_{11}(w)=D_{01}(w)=D_{10}(w)=0$ e $D_{00}(w)=1$ para algum indivíduo w , alguma forma de sinergia está presente. GREENLAND e POOLE (1988) diminuíram esses 16 possíveis tipos de respostas potenciais para 12, considerando que os tipos de respostas devem ser invariantes com relação à sua codificação, ou seja, um tipo de resposta $D_{01}(w)=D_{10}(w)=D_{00}(w)=0$ e $D_{11}(w)=1$ é equivalente a $D_{11}(w)=D_{01}(w)=D_{00}(w)=0$ e $D_{10}(w)=1$ se a resposta X_2 for recodificada por $1-X_2$ (GREENLAND e POOLE 1988). VANDERWEELE e ROBINS (2007b) ressaltaram essa importância de invariância à troca de categorias, mas consideraram essa classificação por tipos de respostas potenciais ainda insuficientes para avaliação de conjuntos suficientes de causas componentes. VANDERWEELE e ROBINS (2007b) propõem uma forma estatística de se testar a sinergia, considerando que é necessário que se assuma que os efeitos são monotônicos, ou seja, a direção dos efeitos de X_1 e X_2 em D é levada em consideração. X_1 tem um efeito monotônico positivo em D se $D_{ij}(w) \geq D_{i'j}(w)$, sempre que $i \geq i'$, e similarmente monotônico para X_2 (o contrário é um efeito monotônico negativo, menor).

A partir dessa ideia de modelo contrafactual, várias condições foram propostas para que modelos estatísticos pudessem ser usados para checar a presença de sinergia entre causas representadas por variáveis binárias (VANDERWEELE e ROBINS 2008, VANDERWEELE 2009, GREENLAND 2009). O modelo estatístico proposto por VANDERWEELE (2009) é um modelo de regressão logística com um fator multiplicativo para o efeito da interação. Essa abordagem já tinha sido empregada anteriormente por HOFFMANN et al. (2005), mas sem uma

abordagem de modelos contrafactuais; os autores apenas usaram a regressão logística com efeito multiplicativo para representar as interações em termos de sinergias.

Um artigo recente (VANDERWEELE et al 2010) sugere o uso de modelos estruturais marginais (MEM) (ROBINS et al. 2000), em vez de regressões lineares tradicionais, para avaliação de presença de sinergia controlando confundimentos. Os MEM são modelos estatísticos baseados na teoria contrafactual, nos quais há a possibilidade de ocorrer uma situação em que X_2 é uma consequência de X_1 , e X_1 pode confundir a relação entre X_2 e D . Por exemplo, o valor esperado de ocorrência da doença pode ser estimado por uma regressão logística da seguinte forma:

$$P[D/X_1=x_1, X_2=x_2, C=c] = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2 + \alpha'_4 c,$$

onde c é um conjunto de variáveis de confusão (α_4 é um vetor de parâmetros). Uma variável de confusão apresenta-se associada a uma variável explicativa e também ao evento e deve ser considerada no modelo, para evitar problemas de efeito nas estimativas das relações entre as variáveis de interesse (HOSMER e LEMESHOW 1989, ALENCAR 2009).

Um MEM equivale a essa regressão logística, se $C=\emptyset$ (por exemplo, quando X_1 e X_2 são aleatorizados). Mas, se há um confundimento, a regressão logística pode não ser suficiente para controlar C . Além disso, pode acontecer de X_1 ter um efeito, L , que afeta X_2 e também D , e o interesse é estimar o efeito conjunto de X_1 e X_2 . Então, considerando L um efeito de X_1 que pode afetar X_2 , temos que o MEM é uma regressão do tipo: $P[Dx_1x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ com uma ponderação. Essa ponderação é produto de duas regressões. Na primeira regressão, calcula-se $P[X_1/C]$ e a segunda $P[X_2/X_1, C, L]$ e o inverso dos valores esperados para a ocorrência da

doença são os pesos v_w^1 e v_w^2 . O peso da regressão final $P[D_{x1x2}]$ é dado por $\text{peso} = v_w^1 \cdot v_w^2$ (VANDERWEELE et al. 2010). Esse tipo de ponderação é bastante parecido com o que propõem ROSENBAUM e RUBIN (1985) e ROSENBAUM (1989) como escore de propensão para transformar um estudo observacional em um estudo quase-aleatorizado. Outros autores sugerem a substituição da regressão logística por modelos aditivos do tipo regressão de chances (em inglês: “linear odds model”), que, no lugar de um efeito multiplicativo, um efeito aditivo seria mais apropriado para a avaliação de presença de sinergia (SKRONDAL 2003).

RICHARDSON e KAUFMAN (2009) apresentaram uma forma simples de se estimar o efeito da interação em modelos log-lineares, mais especificamente no caso da regressão logística (com uso do procedimento `nlmixed` do software SAS). Os autores mostraram como calcular o excesso de risco devido à interação (RERI: relative excess risk due to interaction, também conhecido como ICR: interaction contrast ratio), a partir de um modelo log-linear, mas calculando-se uma estatística de efeitos aditivos (escala de risco relativo) (RICHARDSON e KAUFMAN 2009). HOSMER e LEMESHOW (1992) descreveram uma aproximação do RERI e do seu intervalo de confiança, com base na saída obtida para um ajuste de regressão logística. Posteriormente, ASSMANN et al. (1996) demonstraram que as estimativas do intervalo de confiança com reamostragem via Bootstrap eram melhores. MALDONADO e GREENLAND (2002) também sugeriram que estimativas de Wald para o intervalo de confiança em modelos multiplicativos não apresentavam bom desempenho e sugeriram o uso de estimativas de máxima verossimilhança. RICHARDSON e KAUFMAN (2009) apresentaram a estimativa de $\text{RERI} = \beta_3$, onde β_3 é o parâmetro da interação na regressão logística. O teste para $\beta_3 = 0$ equivale a

testar $RERI=0$, e as estimativas intervalares podem ser calculadas pelo procedimento NLMIXED, que eles propõem no artigo.

SKRONDAL (2003) propõe o cálculo de três estatísticas com base em um modelo linear de chances (em inglês: *linear odds model*): RERI, AP (proporção atribuível devido à interação) e S. A vantagem do cálculo dessas estatísticas por modelos é a possibilidade da inclusão de covariáveis de controle. SKRONDAL(2003) define o modelo de risco da seguinte forma:

$$P(Y=1/x_1, x_2)=R_{jk}=a+b_1x_1+b_2x_2+b_3x_1*x_2, \quad (A.9)$$

onde $x_1=1$ se $j=1$ e $x_1=0$ se $j=0$; $x_2=1$ se $k=1$ e $x_2=0$ se $k=0$.

O modelo de regressão logística é (note-se que os parâmetros são diferentes da equação anterior A.9):

$$R_{jk} = \frac{e^{\alpha+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2}}{1 + e^{\alpha+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2}}. \quad (A.10)$$

Então, a razão de chances é expressa por

$$OR_{jk} = e^{\beta_1x_1} e^{\beta_2x_2} e^{\beta_3x_1x_2}, \quad (A.11)$$

onde o risco relativo será próximo da razão de chances quando o evento é raro e estudos de caso-controle podem utilizar a regressão logística com apenas um ajuste no parâmetro da constante α .

As medidas de interação serão dadas por:

$$RERI = \frac{b_3}{a} = e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1 \quad (A.12)$$

$$AP = \frac{b_3}{a + b_1 + b_2 + b_3} = \frac{e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1}{e^{\beta_1+\beta_2+\beta_3}} \quad (A.13)$$

$$S = \frac{b_1 + b_2 + b_3}{b_1 + b_2} = \frac{e^{\beta_1 + \beta_2 + \beta_3} - 1}{(e^{\beta_1} - 1) + (e^{\beta_2} - 1)} . \quad (\text{A.14})$$

Nesse artigo, SKRONDAL(2003) explica como é feita a inclusão de co-variáveis e conclui que os modelos de risco de chances (A.9) são vantajosos, visto que testar $b_3=0$ equivale a testar sinergia (S), sem necessidade de estatísticas adicionais, como seria necessário para o modelo logístico.

ANEXO B

Cartas do Comitê de Ética em Pesquisa

COMITÊ DE ÉTICA EM PESQUISA – COEP/FSP
Universidade de São Paulo
Faculdade de Saúde Pública

Of.COEP/ 30 /2008

Protocolo	1749
Projeto de Pesquisa	EQUAÇÕES ESTRUTURAIS APLICADAS AO MODELO CAUSAL DE ROTHMAN
Pesquisador(a)	Valéria Troncoso Baltar

26 de FEVEREIRO de 2008.

Prezado(a) Orientador(a),

O Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo - COEP analisou, em sua 1.ª/08 Sessão EXTRAORDINÁRIA, realizada em 22/02/2008, de acordo com os requisitos da Resolução CNS/196/96 e suas complementares, o **protocolo** de pesquisa acima intitulado e o considerou **APROVADO**.

Cabe lembrar que conforme Resolução CNS/196/96 são deveres do (a) pesquisador (a):

1. Comunicar, de imediato, qualquer alteração no projeto e aguardar manifestação deste CEP (Comitê de Ética em Pesquisa), para dar continuidade à pesquisa;
2. Manter sob sua guarda e em local seguro, pelo prazo de 5 (cinco) anos, os dados da pesquisa, contendo fichas individuais e todos os demais documentos recomendados pelo CEP, no caso eventual auditoria;
3. Comunicar, formalmente a este Comitê, quando do encerramento deste projeto;
4. Elaborar e apresentar relatórios parciais e finais;
5. Justificar, perante o CEP, interrupção do projeto ou a não publicação dos resultados.

Atenciosamente,

Claudio Leone
Professor Associado
Coordenador do Comitê de Ética em Pesquisa – FSP/COEP

Ilm. Sr.
Prof. Dr. JULIO CESAR RODRIGUES PEREIRA
DEPARTAMENTO DE EPIDEMIOLOGIA

COMITÊ DE ÉTICA EM PESQUISA – COEP/FSP

Universidade de São Paulo
Faculdade de Saúde Pública

OF.COEP/195/10

11 de agosto de 2010.

Prezada Pesquisadora e Orientador,

O Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo – COEP/FSP, **analisou**, em sua **6ª/10 Sessão Ordinária**, realizada em **6/8/2010**, de acordo com a Resolução n.º 196/96 do Conselho Nacional de Saúde – CNS e suas complementares, o **ADENDO** ao protocolo de pesquisa n.º **1749**, intitulado **"EQUAÇÕES ESTRUTURAIS APLICADAS AO MODELO CAUSAL DE ROTHMAN"**, área temática **GRUPO III**, sob responsabilidade da pesquisadora **Valéria Troncoso Baltar** e orientação do Professor **Júlio César Rodrigues Pereira**, que solicita a substituição da base de dados da pesquisa, considerando-o **APROVADO**.

Cabe lembrar, que conforme Resolução CNS 196/96, são deveres do(a) pesquisador(a): **1. Comunicar**, de imediato, qualquer alteração no projeto e aguardar manifestação deste CEP (Comitê de Ética em Pesquisa), para dar continuidade à pesquisa; **2. Manter sob sua guarda e em local seguro**, pelo prazo de 5 (cinco) anos, os dados da pesquisa, contendo fichas individuais e todos os demais documentos recomendados pelo CEP, no caso eventual auditoria; **3. Comunicar**, formalmente a este Comitê, quando houver o encerramento deste projeto; **4. Elaborar e apresentar relatórios parciais e final**; **5. Justificar**, perante o CEP, interrupção do projeto ou a não publicação dos resultados.

Atenciosamente,

Cláudio Leone
Professor Titular

Coordenador do Comitê de Ética em Pesquisa da FSP/USP

Ilm.º Sr.

Prof. Assoc. **Júlio César Rodrigues Pereira**
Departamento de Epidemiologia
Faculdade de Saúde Pública da USP

ANEXO C

Curriculum Lattes

Julio Cesar Rodrigues Pereira

Bolsista de Produtividade em Pesquisa do CNPq - Nível 2

Julio Cesar Rodrigues Pereira concluiu o doutorado em Epidemiologia pela Faculdade de Saúde Pública da USP em 1995 e Docência Livre em 2004. É médico pela Escola Paulista de Medicina (1977) e Master of Sciences in Medical Sciences pela Universidade de Bristol - UK (1989). Atualmente é Professor Associado da Universidade de São Paulo e Professor Pós-graduação do Instituto Dante Pazzanese de Cardiologia. Publicou 2 livros, 2 capítulos de livro, 54 artigos em periódicos especializados e tem 68 participações em congressos transcritas em anais de eventos ou periódicos. Seu livro "Análise de dados qualitativos" registra 40 citações no ISI e 51 no Scopus e mais de 500 no Google Acadêmico). O ISI registra 38 de seus artigos, 37 citados, num total de 289 citações e H-index = 8. O Scopus registra 34 de seus artigos, 275 citações, H-Index = 9. Participou como convidado em vários eventos no Brasil. Orientou 5 dissertações de mestrado e 4 teses de doutorado nas áreas de Saúde Coletiva, Medicina e Administração. Recebeu 4 prêmios e/ou homenagens. Atua na área de Saúde Coletiva, com ênfase em Epidemiologia. Em suas atividades profissionais interagiu com 144 colaboradores em co-autorias de trabalhos científicos. Em seu currículo Lattes os termos mais frequentes na contextualização da produção científica, tecnológica e artístico-cultural são: Epidemiologia, Avaliação em C&T, Pneumonia, Asma, Modelos matemáticos, Ira, DRI, Infarto Agudo do Miocárdio e Análise de Dados de Survey.

(Texto informado pelo autor)

Última atualização do currículo em 16/12/2010

Endereço para acessar este CV:

<http://lattes.cnpq.br/7576807536889713>

Valéria Troncoso Baltar

Bolsista de Doutorado do CNPq

concluiu graduação em Estatística pela Universidade Estadual de Campinas (1999), especialização em demografia pelo Centro Latinoamericano e Caribenho de Demografia (1999) e mestrado em Estatística pela Universidade de São Paulo (2005). Trabalhou de 2001 a 2007 como estatístico do Laboratório de Epidemiologia e Estatística do Instituto Dante Pazzanese de Cardiologia. Tem experiência na área de Estatística, com ênfase em Análise Multivariada, atuando principalmente nos seguintes temas: migração, cardiologia, saúde pública e ciência e tecnologia. Foi pesquisadora visitante da Imperial College London (St Mary's Campus) e atualmente é aluna de doutorado em Epidemiologia na Faculdade de Saúde Pública da USP.

(Texto informado pelo autor)

Última atualização do currículo em 07/01/2011

Endereço para acessar este CV:

<http://lattes.cnpq.br/9774872677279673>