Supplemental material

Statistical analysis

Concepts and definitions

We extended the conventional twin methods to address issues of censoring at follow-up and competing risk of death. Results would agree with those obtained from the conventional twin approach[1,2] if no censoring or competing risk of death were present. We estimated the genetic influence of lung cancer taking into account three possible outcomes: (1) lung cancer diagnosis, (2) no diagnosis and survival through the end of follow-up, and (3) no diagnosis prior to death from other causes during the follow-up. *Heritability* is defined as the proportion of variance on a scale of disease liability, here for lung cancer that is due to genetic factors in the population. *Casewise concordance* estimates an individual risk of disease conditional on disease in a close relative; in a twin study, it is defined as risk of cancer in a twin, conditional on his/her co-twin having the same cancer. The twin can be as genetically similar as a full sibling (DZ) or identical at the sequence level (MZ). Differences in concordance rates by zygosity provide insights into the influence of genetic *vs.* environmental factors on disease risk under standard assumptions of the twin model.[1]

We defined cohort-specific dates of entry and follow-up. We accounted for left-censoring from variable initiation of cancer registration and right-censoring among those censored at the end of follow-up, censored when lost to follow-up to emigration (<2%), or at competing risk of death by measuring the three possible outcomes for each individual at each time point and modeling the transition from no cancer diagnosis to either diagnosis

or death. Before conducting the pairwise analyses, we examined the individual risk of lung cancer diagnosis by age by estimating cumulative lung cancer incidence using the non-parametric Aalen-Johansen estimator.[3] We modeled potential competing deaths,[4,5] which allows estimation of lung cancer risk in a twin given the occurrence of disease in his/her co-twin. We obtained the case-wise concordances by age. Equality of case wise concordance curves for MZ and DZ pairs by age were tested by Pepe and Mori's test.[5] The overall relative recurrence risks in MZ and DZ pairs were derived from the corresponding concordances and cumulative incidence and the multi locus index was calculated with standard errors obtained by the delta method.[6,7]

Biometric modeling

Quantitative models were analyzed to estimate the magnitude of genetic and environmental influence[1,2] that explains variance in lung cancer liability overall and by smoking status. The genetic and environmental influences on lung cancer liability, particularly on heritability, are estimated for twins with smoking status available, with and without adjustment of prevalence according to levels of smoking status. The role of smoking on heritability of liability to lung cancer is then estimated among pairs in which neither has ever smoked, among pairs where both co-twins are ever (former or current) smokers and among pairs in which both co-twins are current smokers.

Model-fitting

The general approach analyzes lung cancer liability covariance between members of MZ and DZ pairs to decompose variation into additive genetic effects (A), dominant genetic

effects (which model deviations of the heterozygote genotype from the mean of the homozygote genotype) (D), common environmental effects (C), and individually unique environmental effects (E). The genetic parameters of the model are specified on the basis of biologic relationship between the co-twins; MZ twins have the same genomic sequence while DZ twins share on average half of their segregating genes (as do full-siblings).

Within-pair covariance of liability is expressed as κ var(A) + γ var(D) +var(C), where $\kappa = \gamma = 1$ for MZ pairs and $\kappa = 1/2$ and $\gamma = 1/4$ for DZ pairs.[1,2] Due to statistical issues of identifiability, A, D, and C cannot be estimated simultaneously.[2] Therefore, a series of models are tested which allow for sequential testing of the significance of specific parameters. Measurement error is estimated in E as this is the component of variance that does not contribute to within-pair resemblance. Dominance effects are, typically, biologically implausible in the absence of additive effects. The primary models are thus the ACE and ADE models, as well as their sub-models AE, CE and E.

We tested for equal thresholds (i.e., normal quintiles of prevalence) between MZ and DZ twins, which is equivalent to assuming that the risk of disease does not differ by zygosity. The biometric modeling approach we applied is comparable to that of Lichtenstein and colleagues[8] but adjusted for censoring. To test for variation in heritability with age at diagnosis, we estimated the cumulative heritability of lung cancer liability at 60, 70, 80, and 100 years of age. We assessed the fit of the sub-models by the Akaike information criterion.

To correct for possible bias due to censoring at follow-up, individuals were assigned weights obtained by calculating the inverse probability of being censored at time of follow-up. Because censoring is dependent within pairs, the same weight was applied to both twins within a pair.[9] The probabilities of being censored were estimated using the Aalen additive model. We then analyzed the weighted sample of complete observations in order to obtain within-pair dependence estimates corrected for bias and heritability in liability to lung cancer.

The matched case cotwin design using pairs in which one was a smoker and the other was not, providing within pair hazard ratios for the association of smoking with lung cancer diagnosis. This analysis was carried out using a Cox proportional hazards model allowing baseline hazard functions to be specifying for pairs (the stratified Cox model). Given that MZ pairs share the genomic sequence, this provides a direct test of the old hypothesis[10] of shared genes that would underlie both the risk of becoming a smoker and the liability to develop lung cancer.

The statistical program R was used for all analyses with the package *mets*.[11]

References:

1. Neale MC, Cardon LR, North Atlantic Treaty Organization. Scientific Affairs Division. Methodology for genetic studies of twins and families. Dordrecht; Boston: Kluwer Academic Publishers, 1992.

2. Sham P. Statistics in human genetics. London; New York: Arnold; John Wiley & Sons, Inc., 1998.

3. Allignol A, Schumacher M, Beyersmann J. Empirical Transition Matrix of Multi-State Models: The etm Package. J Stat Softw 2011; 38.

4. Scheike TH, Holst KK, Hjelmborg JB. Estimating heritability for cause specific mortality based on twin studies. Lifetime Data Anal 2014; 20:210–33.

5. Scheike TH, Holst KK, Hjelmborg JB. Estimating twin concordance for bivariate competing risks twin data. Stat Med 2014; 33:1193-1204.

6. Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 1990;46:222-8.

7. Risch N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol Biomarkers Prev 2001;10:733-41.

8. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78-85.

9. Van der Laan M, Robins J. Unified methods for censored longitudinal data and causality. Springer, 2003.

10. Fisher RA. Cancer and smoking. Nature 1958;182:596.

11. Holst K, Scheike, T. H. mets: Analysis of Multivariate Event Times, R package version 0.2.8.1, http://lava.r-forge.r-project.org/

Supplemental Table 1. Relative risk of lung cancer in a twin if co-twin is diagnosed to the general risk by age among pairs who were both current smokers at baseline in the NorTwinCan Danish, Finnish, Norwegian, and Swedish cohorts

Age, yrs	**MZ pairs λ (se)**	**DZ pairs λ (se)**	**Multi locus index**
All	3.28 (0.69)	2.45 (0.51)	1.57 (0.73)
40-70	5.52 (2.00)	3.46 (1.27)	1.84 (1.25)
70-80	3.41 (0.84)	2.87 (0.64)	1.29 (0.63)
80-90	3.06 (0.67)	2.58 (0.54)	1.31 (0.62)
90+	3.40 (0.75)	2.30 (0.49)	1.84 (0.90)

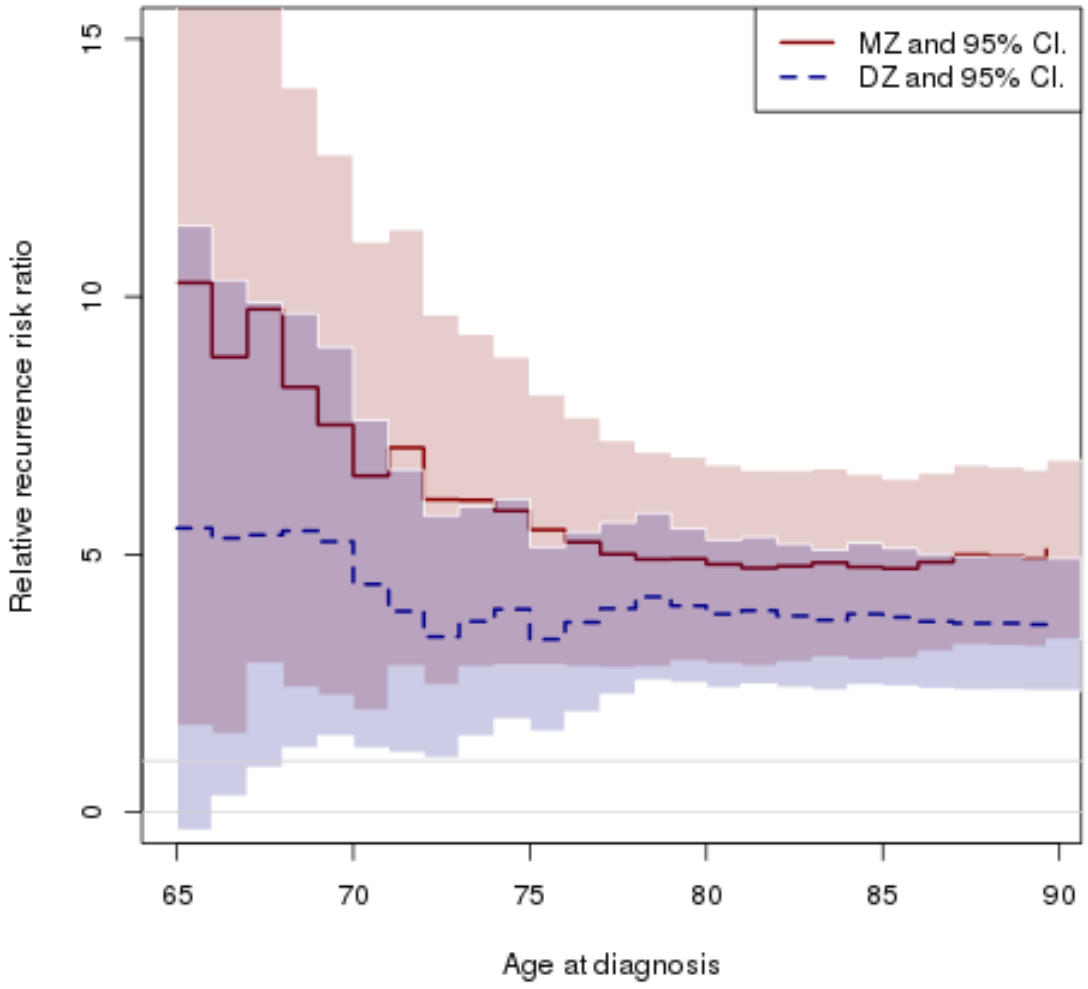
Note: The relative recurrence risk of lung cancer diagnosis, generally referred in genetic epidemiology by **λ.** is in this study obtained as the relative risk of lung cancer in a twin if the co-twin is diagnosed to the general risk of lung cancer in a twin. This is the casewise concordance risk to the cumulative incidence of lung cancer diagnosis and is here estimated for age intervals. The multi-locus index in Table 1 points at an additive genetic effect: The multi-locus index at or below the value two suggests additive contributions of multiple loci and do not indicate epistatic (gene-gene interaction) or dominant effects. The multilocus analysis and the threshold liability model with variance components yield thus consistent results.

Supplemental Figure 1: Relative recurrence risk in MZ and DZ pairs compared to population risk by age in current smokers

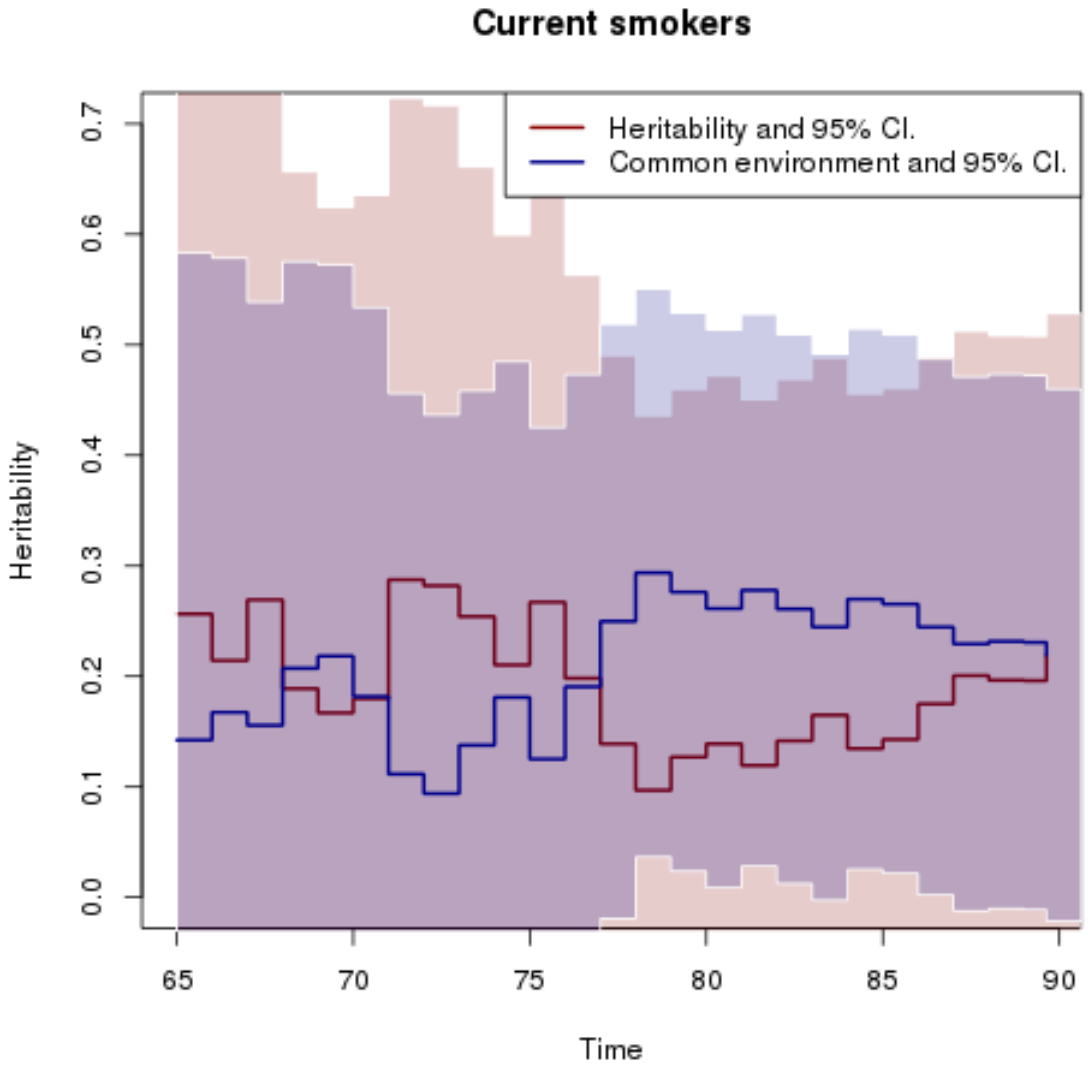
Supplemental Figure 2A: Cumulative heritability of lung cancer (with 95% confidence intervals) and estimated effect due to shared environment by age among current smokers

Supplemental Figure 2B: Cumulative heritability of lung cancer (with 95% confidence intervals) and estimated effect due to shared environment by age among ever smokers

Supplemental Figure 1:



Supplemental Figure 2A:



Supplemental Figure 2B: