Supplementary Information for:

Thorgeirsson et al. "A Variant Associated with Nicotine Dependence, Lung Cancer and Peripheral Arterial Disease**".**

METHODS

Subjects

Smoking Quantity (SQ)

SQ was available from a standardised smoking questionnaire used in deCODE's studies that asks: "How many cigarettes per day do/did you smoke on average (on most days)?" This means that current smokers answer their current consumption and former smokers refer to their consumption in the past. In cases where multiple records were available we recorded the maximum. The SQ was categorised into 4 levels, (1-10, 11-20, 21-30, and 31+ cpd) for two reasons:

i) the data are combined from several questionnaires which used different cut-offs. Some questionnaires ask subjects to give a relatively precise number (i.e. options with a range of 2 cpd, i.e. 1-2, …,13-14, … etc cigarettes per day), and in other cases the questionnaires utilise broader categories, such as those in the FTND. Here, e.g. the values 19-20 cpd and 11-20 cpd are grouped into the same Fagerstrom category, but the majority of individuals answering 11-20 cpd are actually close to 20 cpd. A similar problem occurs in the highest category, which is "more than 45 cpd" for most individuals, but the cut-off can be higher or lower. Thus, overcoming these problems using e.g. group means would leave us with artificial accuracy at the cost of systematic bias.

ii) Most smokers have some variation in their daily consumption and hence it is not reasonable to ask subjects for an exact estimate. This may be particularly true in the case of heavy smokers and introduce noise to the data; the approach based on the FTND combines all individuals smoking more than 30 cpd in one category.

Overall, the FTND categorization simply gives us a convenient method for combining data from different sources, we believe, without much loss of information.

Analysing records with information on smoking in deCODE's phenotypic database, we identified 10,995 ever smokers with information on SQ who had been genotyped for various GWA studies at deCODE and an additional 2,950 individuals with SQ information was individually genotyped.

Further information on smoking status was retrieved from the questions: "Have you ever smoked or used tobacco for as long as a year?" and "Do you currently smoke (i.e. within the last month?" This gives a total of 4,203 genotyped individuals who had never smoked regularly according to questionnaire data, and, 6,388 current and 6,687 former smokers within the group of 13,945 individuals with SQ information.

Nicotine Dependence (ND)

In addition to basic information on smoking behaviour gathered in most projects, several studies specifically address nicotine dependence as part of the phenotypic characterisation. The study of anxiety and depression used the CIDI interview[1] with the full substance disorder modules and the answers on smoking were converted to DSM criteria. Under deCODE's study of ND 3,000 persistent smokers answered detailed questionnaires on smoking behaviour, including the Fagerstrom Test for Nicotine Dependence (FTND)[2] and items probing DSM-IV criteria for nicotine dependence. Individuals diagnosed with other substance disorders were excluded. Among the 2,394 individuals fulfilling criteria for ND there were 121 who fell into SQ level 0, these were treated as cases in the association analyses and hence the number of controls (smokers with an SQ level 0) is 3,506.

Lung Cancer

Iceland

In addition to the characterisation detailed in the main text, the lung cancer patients participating in the genetic study answer a lifestyle questionnaire that includes questions on smoking status (never, former, current), and the quantity and duration of smoking.

The Netherlands. Subjects from three studies are included. The first study investigates causes of urological diseases at the Urology Outpatient Clinic of the RUNMC. The second study is the Nijmegen Biomedical Study (NBS). The third study is part of the Polygene study, an EU 6th framework funded study on genetic variants that modify the risk of cancer.

Urology Outpatient Clinic of the RUNMC: From January 1999 onwards blood samples and life style data were collected from patients visiting the Urology outpatient clinic. These patients were linked to the population-based cancer registry held by the Comprehensive Cancer Centre East (IKO) in Nijmegen. Records were matched on hospital and unique hospital registration number. Among the 7,650 patients with a blood sample available, 29 were diagnosed with lung cancer(26 males; 3 females).

The Nijmegen Biomedical Study (NBS): The NBS is a population-based survey conducted by the Department of Epidemiology and Biostatistics and the department of Clinical Chemistry of the RUNMC. In 2002, 21,756 age and sex stratified randomly selected inhabitants of the municipality of Nijmegen were invited to participate in a study on gene-environment interactions in multifactorial diseases, such as cancer. Participation involved filling out a postal questionnaire including lifestyle and medical history, and to donate an 8.5 ml blood sample. The response to the questionnaire was 43% (N=9,350) and 69% (N=6,468) of the responders donated blood samples. The 6,468 participants for whom a blood sample was available were linked to the cancer registry and matched on date of birth, sex, name, zip code and, if deceased, date of death.

Fifty-one individuals (36 males and 15 females) were recorded in the cancer registry with a diagnosis of lung cancer.

The Polygene study: In this study blood and lifestyle information was collected from patients with prostate cancer (N=957), breast cancer (N=783) and bladder cancer (N=1,022). All of these patients were identified from the IKO regional population-based cancer registry. Only patients of 75 years or younger were included in this study. Fourteen of these patients were also diagnosed with primary lung cancer; 13 males and 1 female.

Spain. Patients were recruited at the Oncology Department of Zaragoza Hospital: All lung cancer cases and 865 of the 1507 control individuals answered a lifestyle questionnaire, including questions on smoking status (never, former, current), and the amount of smoking.

Genotyping

All 10,995 samples in the genome-wide association study of smoking quantity were genotyped using genotyping systems and specialised software from Illumina (Human Hap300 and Human Hap300-duo+ Bead Arrays, Illumina)[3]. In total, 311,388 single-nucleotide polymorphism (SNP) markers, distributed across the human genome, were common to both platforms. For the association analysis, we used 306,207 SNP markers because 5,181 were deemed unusable due to low yield, deviations from Hardy-Weinberg expectations, or discrepancies in genotype frequencies between the two arrays. Samples with a call rate below 98% were excluded prior to analysis. For subsequent studies further genotypes for rs1051730 were retrieved from the Illumina data available in deCODE's genotype database, and 8,566 Icelandic samples and the foreign study groups were genotyped using Centaurus (Nanogen) for rs1051730. A total of 1,879 individuals were genotyped with both technologies, four (0.2%) mismatches were observed and the respective individuals were excluded from the analysis.

Statistical Analysis

Adjustment for relatedness in the Icelandic studies

Evaluation of statistical significance took the relatedness of the Icelandic individuals into account by dividing the test-statistic with a correction factor. For the GWA this was done by the method of genomic control[4] using all 306,207 SNPs passing quality control. In all other comparisons genotype information for the total number of tested individuals was only available for SNP rs1051730, and the correction factor for the $\chi 2$ test-statistic was determined applying a simulation procedure using the known genealogy which we had previously employed[5]. We simulated 100,000 sets of genotypes for the SNP through the Icelandic genealogy of 739,000 individuals. The simulated genotypes were used in the applied tests resulting in 100,000 tests under the null hypothesis and the mean of the respective $\chi 2$ test-statistics gives the correction factor.

The correction factor for the GWA derived by genomic control and simulation were 1.15 and 1.11, respectively. The other correction factors in the smoking-related phenotypes calculated by simulation were: 1.18 for the extended SQ study group of 13,945 smokers, 1.15 and 1.09 for the comparison of the 4,203 never smokers *vs*. all 13,945 smokers and 3,627 low quantity smokers, respectively, 1.21 and 1.19 for testing the 2,394 individuals with ND *vs*. the remaining population controls and *vs*. the low quantity smokers, respectively, 1.07 for the comparison of the 6,388 current smokers *vs*. 6,687 ex-smokers. For the LC and the PAD study the correction factors were 1.04 and 1.06, respectively.

Regression Analysis

The year of birth had been rounded to the nearest 5. When year of birth adjustment was applied to study the effect of the variant, year of birth was treated as a categorical variable with four levels: ≤ 1930, 1935 to 1945, 1950 to 1960, and ≥ 1965. This divided the 13,945 smokers studied

in groups of 3774, 3416, 4027 and 2728, which was the closest we could get to having four groups of equal size. The same categories were applied when analysing the data from Spain and the Netherlands.

Genotypic Odds Ratios

In general, the odds ratios for rs1051730 were calculated assuming a multiplicative model, i.e. the risks of the two alleles a person carries are expected to multiply. For example, if OR is the risk of T relative to C, then the risk of a homozygote TT individual will be OR times that of a heterozygote CT, and OR^2 times that of a homozygote CC. Additionally, genotypic ORs were calculated under the assumption of Hardy-Weinberg equilibrium in the controls (no control population showed a deviation from Hardy-Weinberg equilibrium).

RESULTS

Genome-wide association of Smoking Quantity

SNPs reaching genome-wide significance in the GWA study of SQ level.

Allele T of rs1051730 in the CHRNA3 gene was most strongly associated to SQ, and another six SNPs, all of them located on chromosome 15q24, passed the threshold of genome-wide significance ($P < 2 \times 10^{-7}$), The SNPs, their physical position, r^2 to rs1051730, their p-values from the scan and after adjustment for rs1051730 are displayed in Supplementary Table 1.

Nicotine Dependence and Current Smoking Status

Allele frequency of the variant according to FTND and DSM IV-items

Supplementary Table 2 shows the distribution of the items from our main ND tools. Our study primarily recruited smokers who had at least smoked 15 cigarettes per day at one point in their lives, to enable us to collect a large sample of individuals with ND according to these criteria.

Due to the specific recruitment of the individuals only the ones who obtain a score that meets ND criteria and have complete records are shown. Thus, the data cannot reflect the distribution of the answers of all individuals who ever smoked on these items. Since these data were not collected population-based, a quantitative analysis is not very meaningful and the true influence of the variant on the single items could substantially increase when more low-quantity smokers were included. The Table displays the results based on 1,364 and 1,324 individuals who have complete records for DSM IV criteria and the FTND, respectively. A total of 979 individuals are included in both data sets.

Estimating the effect of the variant on ability to quit by analysing current and former smokers

The current smoker status (based on the question "Do you currently smoke (i.e. within the last month)?") is available for 13,075 individuals with SQ information and is used as dependent variable in a logistic regression model. Supplementary Table 3 shows the effect of the variant adjusted for sex and year of birth (same categories as in the initial SQ analysis). The risk is higher in females, but the interaction term is not significant (P=0.10). The effect is similar when corrected for SQ (OR=1.06, 95% CI:1.00–1.12, P=0.036).

Lung Cancer and Peripheral Arterial Disease

Estimating the effect of the variant on LC and PAD through its effect on smoking quantity

Here we investigate what the effect of the variant on LC would be under the simplifying assumption that the LC risk is only due to effect of the variant on SQ level as measured. In the 13,945 Icelandic smokers studied, 501 are known lung cancer cases. Using logistic regression adjusted for sex and year of birth, compared to SQ level 0, levels 1 to 3 are estimated to have relative risks of 2.1, 2.4 and 2.9 for lung cancer, respectively. These relative risk estimates are

not inconsistent with numbers reported in other studies[6, 7]. Notably, the biggest jump in relative risks occurred between level 0 and level 1. Assuming these relative risk estimates and applying them to the distribution of smokers in various SQ levels as displayed in Table 3, frequency of the variant in lung cancer patients can be calculated as a weighted average. Specifically, the predicted frequency is

$$[(0.305\times0.260)+(0.350\times0.459\times2.1)+(0.380\times0.214\times2.4)+(0.391\times0.067\times2.9)]$$

divided by

$$[0.260+(0.459\times2.1)+(0.214\times2.4)+(0.067\times2.9)],$$

or 35.6%. Note that this calculation assumes only smokers would have lung cancer (non-smokers are given weight zero) and hence could over-estimate the frequency of the variant. Still, compared to the frequency of the variant in population controls (34.4%), the OR is only around 1.05. Note that since the frequency of the variant in SQ level 1 is only 35%, to increase the predicted frequency requires increasing the weights of SQ levels 2 and 3. However, even if we doubled the relative risks for SQ levels 2 and 3, from 2.4 and 2.9 to 4.8 and 5.8 respectively, the frequency and OR predicted for lung cancer patients would only increase respectively to 36.3% and 1.09.

Doing the same calculations for PAD leads to a similar conclusion. Compared to SQ level 0, levels 1 to 3 show homogeneous relative risks of 1.56, 1.52 and 1.57 for PAD, respectively. The increase in relative risks occurs between level 0 and level 1. Plugging these values in the formula above, the expected frequency in PAD patients is 35.2%. Again this calculation assumes only smokers would have PAD and compared to the frequency of the variant in population controls (34.4%), the OR is around 1.04.

Frequency of allele T of rs1051730 and histological types of lung cancer.

Supplementary Table 4 gives the frequency of the risk allele in the histological types of lung cancer in the three study groups. In Iceland the frequency of the variant is the highest in small cell lung carcinoma (44.3%) followed by adenocarcinoma (41.8%), the two histological types most strongly associated with smoking. The frequency of the variant is lowest in the group "other specified histologies" where it is close to the control frequency (33.7%). This category includes carcinoid tumors which are known not to be associated with smoking.

Genotypic ORs

Genotypic ORs for our main findings are displayed in Supplementary Table 5. The test of the multiplicative model versus the full model is not significant for all three phenotypes and the obtained genotypic ORs are very close to the ones inferred from the allelic ORs. Hence we have no indication that the underlying genetic model is not multiplicative.

Sex-specific results

Among LC cases the fraction of females is higher in Iceland than in Spain or the Netherlands, which is in agreement with the incidence of LC in the respective countries[8]. The sex-specific results for the variant in LC, PAD and ND are displayed in Supplementary Table 6. The OR estimates vary between the sexes and higher risks are observed in males for PAD and ND, whereas females have higher risks for LC. To address the question, whether these differences are significant, we tested directly male cases versus female cases. The resulting ORs are: LC (female vs. male): OR=1.15, 95% CI:0.94–1.40, P=0.16; PAD (male vs. female): OR=1.11, 95% CI:0.99–1.24, P=0.071; ND (male vs. female): OR=1.13, 95% CI:1.00–1.28, P=0.052. Even though all comparisons are not significant, further studies might substantiate the observed trends.

ACKNOWLEDGEMENTS

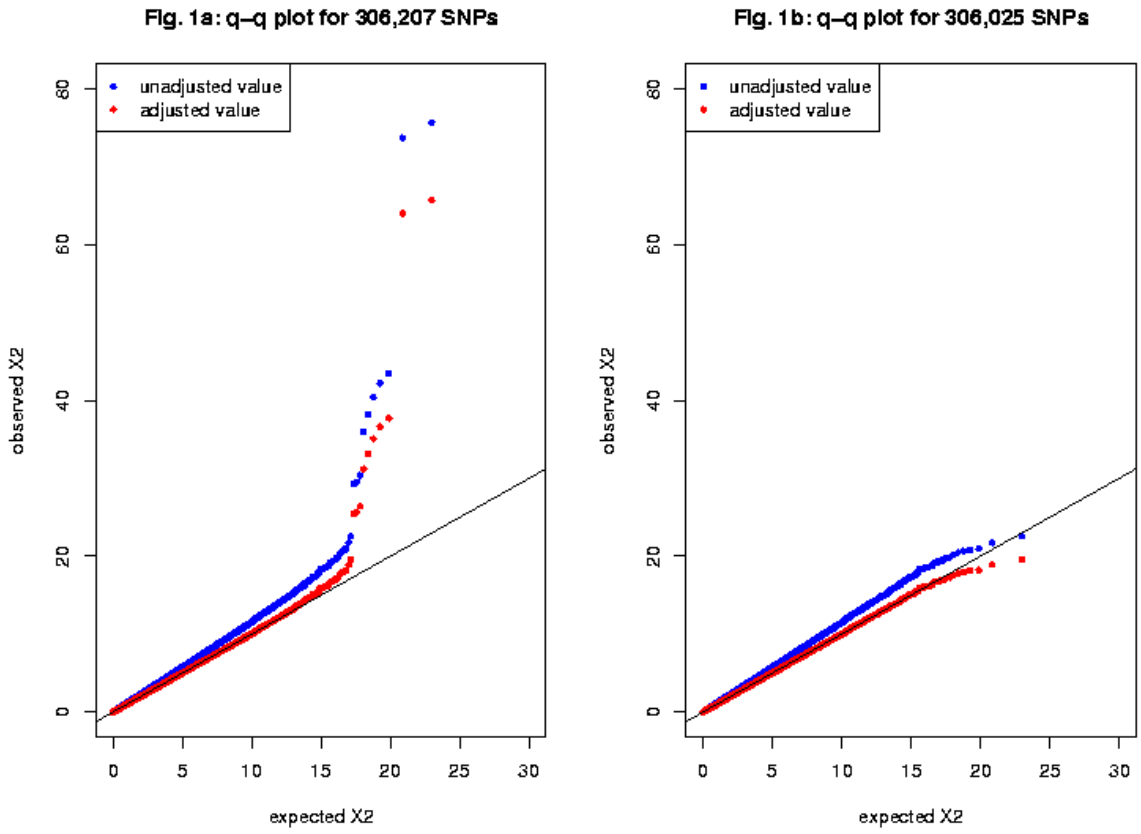
This work was supported in part by grants from the NIH (R01-DA017932) and the European Commission (GENADDICT: LSHM-CT-2004-005166).

REFERENCES

1. Thorgeirsson, T. E. et al. Anxiety with panic disorder linked to chromosome 9q in Iceland. Am J Hum Genet 72, 1221-30 (2003).
2. Heatherton, T. F., Kozlowski, L. T., Frecker, R. C. & Fagerstrom, K. O. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. Br J Addict 86, 1119-27 (1991).
3. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. Nat Genet 38, 659-62 (2006).
4. Devlin, B., Bacanu, S. A. & Roeder, K. Genomic Control to the extreme. Nat Genet 36, 1129-30; author reply 1131 (2004).
5. Grant, S. F. et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38, 320-3 (2006).
6. Haiman, C. A. et al. Ethnic and racial differences in the smoking-related risk of lung cancer. N Engl J Med 354, 333-42 (2006).
7. Stellman, S. D. et al. Lung cancer risk in white and black Americans. Ann Epidemiol 13, 294-302 (2003).
8. Tyczynski, J. E., Bray, F. & Parkin, D. M. Lung cancer in Europe in 2000: epidemiology, prevention, and early detection. Lancet Oncol 4, 45-55 (2003).

Supplementary Figure 1a: Quantile-Quantile plot of the 306,207 chi-square statistics (X2) from the GWA study on the quantitative trait SQ level. The results are for 10,995 individuals from the Icelandic population. The unadjusted statistics are displayed in blue, the statistics divided by a scaling factor of 1.15, which was derived by genomic control, are displayed in red.

Supplementary Figure 1b: Quantile-Quantile plot of 306,025 chi-square statistics (X2) from the GWA study after removing 182 markers located within 1 Mb of rs1051730. The unadjusted statistics are displayed in blue, the scaling factor is again 1.15 and the adjusted statistics are displayed in red.



Supplementary Table 1: Genome-wide significant SNPs on chromosome 15 and their results

SNP	position bp (NCBI build 36)	r^2 to rs1051730	P	P adjusted for rs1051730
rs1051730	76681394	-	5×10^{-16}	-
rs8034191	76593078	0.93	1×10^{-15}	0.60
rs4887077	76765419	0.43	8×10^{-10}	0.30
rs11638372	76770614	0.44	1×10^{-9}	0.35
rs1996371	76743861	0.45	3×10^{-9}	0.55
rs6495314	76747584	0.44	8×10^{-9}	0.64
rs2036534	76614003	0.14	2×10^{-8}	4×10^{-3}

Supplementary Table 2a: Frequency of rs1051730 risk allele T according to items from the DSM IV questionnaire (1,364 individuals, 67% females, mean age 47.7 years). The overall mean of the variant in the group is 38.8%.

Item	Yes		No	
	N	freq.	N	Freq
Tolerance	375	0.404	989	0.382
Withdrawal	1039	0.397	325	0.357
Using more than intended	1216	0.385	148	0.412
Persistent desire to quit	1310	0.391	54	0.306
Time spent on substance	894	0.393	470	0.378
Giving up activities	230	0.426	1134	0.380
Despite health consequences	940	0.392	424	0.379
Total number of criteria fulfilled				
3 criteria	379	0.364		
4 criteria	377	0.387		
5 criteria	356	0.376		
6 criteria	185	0.443		
7 criteria	67	0.433		

Supplementary Table 2b: Frequency of rs1051730 risk allele T according to items (points in parentheses) from the FTND questionnaire (1,324 individuals, 64% females, mean age 47.9 years). The overall mean of the variant in the group is 38.7%.

Item	Yes		No	
	N	freq.	N	freq.
How soon after you wake up do you smoke your first cigarette? (max. 3)				
after 60 minutes (0)	7	0.214		
31-60 minutes (1)	149	0.352		
6-30 minutes (2)	905	0.393		
within 5 minutes (3)	263	0.394		
Do you find it difficult to refrain from smoking in places where it is forbidden? (1)	347	0.372	977	0.393
Which cigarette would you hate most to give up, first one in the morning? (1)	1137	0.394	187	0.350
How many cigarettes per day do you smoke? (1)				
SQ level 0	40	0.325		
SQ level 1	725	0.377		
SQ level 2	434	0.392		
SQ level 3	125	0.456		
Do you smoke more frequently during the first hours after waking up? (1)	393	0.401	931	0.382
Do you smoke if you are so ill that you are in bed most of the day? (1)	782	0.405	542	0.363
Total number of points				
4 points	370	0.349		
5 points	342	0.399		
6 points	288	0.387		
7 points	184	0.383		
8 points	98	0.464		
9 points	35	0.471		
10 points	7	0.500		

Supplementary Table 3: Multiple logistic regression model for current smoker status.

Variable	Estimate (95%CI)	P
Copies of allele T	1.07 (1.01 – 1.13)	0.015
Male sex	0.80 (0.74 – 0.86)	$< 1 \times 10^{-8}$
Categorical year of birth	-	$< 2 \times 10^{-16}$

Supplementary Table 4: Frequency of rs1051730 risk allele T in histological subtypes

Population	Iceland 30184 controls, freq. 34.4%		Spain 1474 controls, freq. 39.0%		Netherlands 2018 controls, freq. 31.4%	
Histology	**N**	**freq.**	**N**	**freq.**	**N**	**freq.**
Adenocarcinoma	275	41.8%	54	51.9%	26	42.3%
Squamous cell carcinoma	149	34.9%	105	50.0%	42	35.7%
Small cell carcinoma	87	44.3%	55	40.0%	10	20.0%
Carcinoma NOS	60	40.0%	35	55.7%	1	0.0%
Large cell carcinoma	30	40.0%	18	44.4%	4	37.5%
Other (incl. Carcinoid)	46	33.7%	2	0.0%	7	28.6%

Supplementary Table 5. Model-free estimates of the genotypic odds ratio of rs1051730 for the main findings.

		Genotypic OR[a]			
Study	**Allelic OR**	**CC**	**CT (95% CI)**	**TT (95% CI)**	***P*[b]**	***P*[c]**
Lung cancer						
Iceland	1.27	1	1.23 (1.04 - 1.46)	1.64 (1.29 - 2.09)	0.66	$2.0x10^{-4}$
Foreign combined	1.38	1	1.50 (1.18 - 1.92)	1.87 (1.34 - 2.62)	0.38	$2.4x10^{-4}$
All combined	1.31	1	1.32 (1.15 - 1.52)	1.70 (1.40 - 2.08)	0.89	$1.1x10^{-7}$
PAD						
Iceland	1.18	1	1.14 (1.02 - 1.28)	1.40 (1.19 - 1.66)	0.53	$2.3x10^{-4}$
Foreign combined	1.23	1	1.30 (1.12 - 1.51)	1.46 (1.14 - 1.87)	0.24	$1.3x10^{-3}$
All combined	1.19	1	1.21 (1.10 - 1.32)	1.41 (1.23 -1.62)	0.76	$8.9x10^{-7}$
ND						
Iceland only	1.40	1	1.39 (1.25 - 1.55)	1.97 (1.65 -2.35)	0.89	$6.7x10^{-14}$

[a] Genotype odds ratios for heterozygous- (CT) and homozygous carriers (TT) compared with non-carriers (CC).

[b] Test of the multiplicative model (the null hypothesis) versus the full model (one degree of freedom).

[c] Test of no effect (the null hypothesis) versus the full model (two degrees of freedom).

OR - odds ratio, CI - confidence interval.

Supplementary Table 6: Sex-specific odds ratio analysis of rs1051730 allele T for main findings in LC, PAD and ND.

Study Group	controls		males				females			
	n	freq	n	OR	(95% CI)	*P*	n	OR	(95% CI)	*P*
Lung cancer										
Iceland	28,752	0.342	346	1.12	(0.96 - 1.32)	0.15	319	1.45	(1.23 - 1.70)	8.3×10^{-6}
Spain	1,474	0.390	238	1.50	(1.24 - 1.82)	4.4×10^{-5}	31	1.21	(0.73 - 2.01)	0.47
The Netherlands	2,018	0.314	71	1.30	(0.92 - 1.85)	0.14	19	0.78	(0.38 - 1.59)	0.50
foreign combined	3,492	-	309	1.45	(1.22 - 1.72)	1.8×10^{-5}	50	1.04	(0.69 - 1.57)	0.86
all combined	32,244	-	655	1.26	(1.13 - 1.42)	7.4×10^{-5}	369	1.38	(1.19 - 1.61)	2.6×10^{-5}
PAD										
Iceland	28,752	0.342	926	1.23	(1.11 - 1.35)	5×10^{-5}	577	1.10	(0.97 - 1.24)	0.15
New Zealand	435	0.274	252	1.44	(1.13 - 1.82)	0.0027	189	1.24	(0.95 - 1.61)	0.12
Austria	403	0.352	322	1.28	(1.03 - 1.58)	0.025	135	1.03	(0.77 - 1.37)	0.84
Sweden	140	0.304	92	0.95	(0.63 - 1.43)	0.82	80	1.38	(0.91 - 2.07)	0.13
Italy	234	0.378	111	1.19	(0.86 - 1.64)	0.31	54	1.09	(0.71 - 1.67)	0.70
foreign combined	1,212	-	777	1.27	(1.11 - 1.45)	5.3×10^{-4}	458	1.16	(0.99 - 1.37)	0.067
all combined	29,964	-	1,703	1.24	(1.15 - 1.34)	1.0×10^{-7}	1,035	1.12	(1.02 - 1.24)	0.024
ND Iceland										
ND vs. 1-10 cpd	3,506	0.303	800	1.52	(1.34 - 1.72)	3×10^{-11}	1,594	1.34	(1.22 - 1.48)	2.2×10^{-9}