Appendix A. Tobacco Use Status and Smoking Attribution of Incident Cancer Cases

Methods

We examined tobacco use status among incident cancer cases, as recorded in the Massachusetts Cancer Registry (MCR) case files. We focused primarily on 11 cancer types that the U.S. Surgeon General[1] and the International Agency for Research on Cancer (IARC)[2] consider tobacco-related. These included cancers of the bronchus and lung, cervix uteri, colon and rectum, esophagus, kidney and renal pelvis, larynx, liver and intrahepatic bile ducts, oral cavity and pharynx, pancreas, stomach, and urinary bladder. There was an insufficient number of acute myeloid leukemia cases to include in this analysis.

The MCR tobacco use variable includes the following categories: never used tobacco, current cigarette smoker, current cigar or pipe smoker, current snuff/chew/smokeless tobacco user, current combination user, previous tobacco use, and unknown. Current cigar or pipe smoking was rare (N=1-4 cases); these were combined with current cigarette smokers under the category of "current smoking." Current snuff/chew/smokeless tobacco use was also rare (N=1-4 cases), but these were not combined with current smokers since the associations between smokeless tobacco use and cancer appear to differ from those reported for smoked tobacco.[3] There were no cases with current combination tobacco use. Given the paucity of smokeless tobacco use among cases, we classified all instances of previous tobacco use as "former smoking."

In computing the proportions of current and former smokers among incident cases, we excluded cases with "unknown" tobacco use status from the denominator. The frequency of unknown tobacco use status was 20% for all incident cancers and 17% across the 11 tobacco-related cancer types. Within these 11 cancer types, the proportion with unknown tobacco use status ranged from ≤10% for 3 cancer types to 36% for liver and intrahepatic bile duct cancer. About 12% of bronchus and lung cancer cases had unknown tobacco use status. In a sensitivity analysis described below, we used imputation methods to reclassify unknown tobacco status to one of the other tobacco use categories based on other variables in the dataset.

For the 11 tobacco-related cancer types, we estimated the proportion of incident cases attributable to tobacco smoking using the following population attributable fraction (PAF) formula.[4-6]

$$PAF = \frac{p_c(RR_c - 1)}{RR_c} + \frac{p_f(RR_f - 1)}{RR_f}$$

where p_c and p_f are the proportions of cases occurring in current and former smokers, and RR_c and RR_f are the relative risks of a given cancer type among current and former smokers as compared to never smokers. The PAF represents the proportion of cancer cases that would not have occurred in the absence of tobacco smoking.[6] PAF estimation is the principal methodology used by the CDC,[7] the U.S. Surgeon General,[1] and the Global Burden of Disease investigators[8,9] to quantify the population health burden attributable to tobacco use and other risk behaviors.

In our primary analyses, we computed PAFs using p_c and p_f estimates that were based on the denominator of cases in the study cohort with known tobacco use status. In a sensitivity analysis, we used the discriminant function method[10] to replace unknown tobacco use status with an

imputed tobacco use value based on the age at diagnosis, sex, race, and cancer type of the case patient.

Our PAF calculations incorporated relative risks (RRs) from published meta-analyses examining the relationship between tobacco smoking and selected cancers. For most cancer types, we used RR estimates from a 2008 meta-analysis[11] of studies identified by a 2004 IARC monograph on the carcinogenicity of tobacco smoking,[12] except when a more recent or comprehensive meta-analysis was available.[13-16] We used RRs specific to North America when there was evidence of geographic heterogeneity.[13,14] We used RRs specific to cancer incidence when these were presented separately from those for mortality.[14] In keeping with MCR standards,[17] we analyzed oral cavity and pharynx cancers as a single category, although meta-analytic findings have suggested a difference in the RRs for oral cavity cancer and pharynx cancer when considered separately.[11] We conservatively used the lower of the two RR estimates (oral cavity cancer) in calculating the PAF for the composite category.

Since PAF computation relies on tobacco use data and RR estimates with an inherent degree of imprecision, we used Monte Carlo simulations to account for this uncertainty and to generate 95% CIs for our PAFs.[6,9,18] We took 1,000 random draws from the log-normal distribution of RR estimates and 1,000 random draws from the binomial distributions of current and former smoking among incident cases to generate 1,000 PAFs for each cancer type. We ordered the 1,000 simulated PAFs and reported the 25th, 500th, and 975th values as the lower confidence bound, point estimate, and upper confidence bound, respectively. We then multiplied the PAF for each cancer type by the corresponding count of incident cases to estimate the number of

smoking-attributable cases for each cancer type. We summed this product across all 11 cancer types to generate a composite estimate of the number of incident cancer cases attributable to tobacco smoking.

Results

Among incident cases where tobacco use status was known, 88% of bronchus and lung cancers, 83% of oral cavity and pharynx cancers, 74% of liver and intrahepatic bile duct cancers, and 63% of colon and rectum cancers occurred in current smokers. Across 248 incident cases of all 11 tobacco-related cancer types where tobacco use status was known, 75% occurred in current smokers and 15% occurred in former smokers. By comparison, in the 121 incident cases of non-tobacco related cancers where tobacco use status was known, 60% occurred in current smokers and 17% occurred in former smokers.

The simulated PAFs for each cancer type are shown in the Table. PAFs calculated with imputed tobacco use values in place of unknown values were not substantively different from those based only on cases with known tobacco status. We focused primarily on the PAFs from non-imputed analyses since they were slightly more conservative. Based on these estimates, approximately 88% (95% CI 81%-91%) of bronchus and lung cancer cases, 72% (95% CI 36%-88%) of larynx cancer cases, 60% (95% CI 45%-71%) of oral cavity and pharynx cancer cases, and 56% (95% CI 42%-67%) of urinary bladder cancer cases were smoking attributable. Across all 11 tobacco-related cancer types, about 157 (95% CI 147-166) cases were smoking-attributable. This represents 34% (95% CI 32%-36%) of all incident cancer cases in the study cohort.

Comparison to Other Settings

In the U.S. general population, about 82% of lung cancer deaths,[1] 67% of oral cavity and pharynx cancer deaths,[19] and 29% of all cancer deaths[20] are attributable to cigarette smoking. In the United Kingdom, an estimated 86% of incident lung cancer cases, 65% of incident oral cavity and pharynx cancer cases, and 19% of all incident cancer cases are tobacco-attributable.[21] In comparison to these general population estimates, the higher proportion (34%) of all incident cancer cases attributable to tobacco smoking in the BHCHP cohort likely reflects both the excess rates of certain smoking-attributable cancers (e.g. bronchus and lung cancer and oral cavity and pharynx cancer) as well as the significantly lower rates of certain non-tobacco-related cancers (e.g., prostate cancer and female breast cancer).

Appendix A Table. Tobacco Population Attributable Fractions (PAFs) by Cancer Site/Type

Cancer site/type	Data source for relative risk	PAF (95% CI), non-imputed[a]	PAF (95% CI), imputed[b]
Bronchus and lung	Lee et al.[13] [c]	0.88 (0.81-0.91)	0.88 (0.81-0.91)
Cervix uteri	Gandini et al.[11]	0.20 (0.05-0.36)	0.22 (0.09-0.37)
Colon and rectum	Botteri et al.[14] [d]	0.14 (0.07-0.21)	0.14 (0.07-0.22)
Esophagus	Gandini et al.[11]	0.49 (0.29-0.61)	0.50 (0.32-0.62)
Kidney and renal pelvis	Gandini et al.[11]	0.26 (0.14-0.37)	0.28 (0.17-0.39)
Larynx	Gandini et al.[11]	0.72 (0.36-0.88)	0.73 (0.43-0.88)
Liver and intrahepatic bile ducts	Lee et al.[15]	0.26 (0.16-0.34)	0.26 (0.16-0.35)
Oral cavity and pharynx	Gandini et al.[11] [e]	0.60 (0.45-0.71)	0.60 (0.46-0.70)
Pancreas	Iodice et al.[16]	0.21 (0.10-0.32)	0.22 (0.11-0.32)
Stomach	Gandini et al.[11]	0.18 (0.05-0.31)	0.18 (0.07-0.31)
Urinary bladder	Gandini et al.[11]	0.56 (0.42-0.67)	0.58 (0.45-0.68)

PAF, population attributable fraction

[a] PAF estimates based on current and former smoking proportions among those with known tobacco status at the time of diagnosis

[b] PAF estimates based current and former smoking proportions after imputing tobacco status for unknown cases using discriminant function methods

[c] North American studies, any tobacco product, random effects RR estimates

[d] North American studies, incidence RR estimates

[e] Oral cavity RR estimates

Appendix A References

1. U.S. DHHS. The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General. Atlanta, GA: U.S. DHHS, CDC, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2014.
2. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Tobacco Smoking. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 100E: Personal Habits and Indoor Combustions*. Lyon: International Agency for Research on Cancer; 2012:43-211.
3. Lee PN, Hamling J. Systematic review of the relation between smokeless tobacco and cancer in Europe and North America. *BMC Med.* 2009;7:36. http://dx.doi.org/10.1186/1741-7015-7-36.
4. Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol.* 1974;99(5):325-332.
5. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott; 1998.
6. Steenland K, Armstrong B. An overview of methods for calculating the burden of disease due to specific risk factors. *Epidemiology.* 2006;17(5):512-519. http://dx.doi.org/10.1097/01.ede.0000229155.05644.43.
7. CDC. Methodology: Smoking-Attributable Mortality, Morbidity, and Economic Costs (SAMMEC). http://apps.nccd.cdc.gov/sammec/methodology.asp. Accessed January 22, 2014.
8. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010:

a systematic analysis for the Global Burden of Disease Study 2010. *Lancet.* 2012;380(9859):2224-2260. http://dx.doi.org/10.1016/S0140-6736(12)61766-8.

9. U.S. Burden of Disease Collaborators. The state of U.S. health, 1990-2010: burden of diseases, injuries, and risk factors. *JAMA.* 2013;310(6):591-608. http://dx.doi.org/10.1001/jama.2013.13805.

10. Allison PD. Paper 113-30: Imputation of Categorical Variables with PROC MI. *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.; 2005.

11. Gandini S, Botteri E, Iodice S, et al. Tobacco smoking and cancer: a meta-analysis. *Int J Cancer.* 2008;122(1):155-164. http://dx.doi.org/10.1002/ijc.23033.

12. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. *IARC Monographs on the Evaluation of Circinogenic Risks to Humans, Volume 83: Tobacco Smoke and Involuntary Smoking*. Lyon: International Agency for Research on Cancer; 2004.

13. Lee PN, Forey BA, Coombs KJ. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer.* 2012;12:385. http://dx.doi.org/10.1186/1471-2407-12-385.

14. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal cancer: a meta-analysis. *JAMA.* 2008;300(23):2765-2778. http://dx.doi.org/10.1001/jama.2008.839.

15. Lee YC, Cohet C, Yang YC, Stayner L, Hashibe M, Straif K. Meta-analysis of epidemiologic studies on cigarette smoking and liver cancer. *Int J Epidemiol.* 2009;38(6):1497-1511. http://dx.doi.org/10.1093/ije/dyp280.

16. Iodice S, Gandini S, Maisonneuve P, Lowenfels AB. Tobacco and the risk of pancreatic cancer: a review and meta-analysis. *Langenbecks Arch Surg.* 2008;393(4):535-545. http://dx.doi.org/10.1007/s00423-007-0266-2.

17. Gershman ST, Massachusetts Cancer Registry. Cancer Incidence and Mortality in Massachusetts 2004-2008: Statewide Report: Massachusetts Department of Public Health, Bureau of Health Information, Statistics, Research, and Evaluation; August 2011.

18. Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int J Epidemiol.* 2004;33(6):1389-1397. http://dx.doi.org/10.1093/ije/dyh276.

19. Office of the Surgeon General, Office on Smoking and Health. The Health Consequences of Smoking: A Report of the Surgeon General. Atlanta, GA: CDC; 2004.

20. Jacobs EJ, Newton CC, Carter BD, et al. What proportion of cancer deaths in the contemporary United States is attributable to cigarette smoking? *Ann Epidemiol.* 2014;25(3)179-182. http://dx.doi.org/10.1016/j.annepidem.2014.11.008.

21. Parkin DM. 2. Tobacco-attributable cancer burden in the UK in 2010. *Br J Cancer.* 2011;105(Suppl 2):S6-S13. http://dx.doi.org/10.1038/bjc.2011.475.

Appendix B. Methods Used to Correct for Discrepancies Between Death Certificates and Cancer Registry Records

Estimating the Correction Factor

As described in the manuscript, we revised the number of deaths due to certain cancer types in the homeless cohort when the reported cause of cancer death on the death certificate differed from the cancer type listed in the Massachusetts Cancer Registry (MCR) record. However, we did not have the ability to directly identify and resolve similar discrepancies that may have occurred in the general population because cancer mortality rates in Massachusetts were obtained from the CDC Wide-ranging Online Data for Epidemiologic Research (WONDER) database[1] while cancer incidence rates in Massachusetts were obtained from a published MCR report,[2] and both sources provided only aggregate data. In order to avoid biasing the standardized mortality ratio (SMR) estimates by dividing a corrected numerator (observed cancer deaths in the homeless cohort) by an uncorrected denominator (expected cancer deaths based on Massachusetts mortality data), we adjusted the SMR denominator using a correction factor.[3,4] We calculated the correction factors for each cancer type based on data from a study examining discrepancies between cancer death records and cancer registry records for 265,863 cancer decedents in three states.[5] This study compared the cancer type listed on the death certificates of cancer decedents with the cancer type listed in the corresponding cancer registry record and calculated confirmation rates (positive predictive values) and detection rates (sensitivities) for death certificates in identifying deaths due to various cancer types. We calculated the correction factor for each cancer type by dividing the death certificate confirmation rate by the death

certificate detection rate.[3] The death certificate confirmation rate, detection rate, and correction factor for identifying deaths due to cancer *x* can be mathematically displayed as follows:

$$Confirmation\ rate = \text{positive predictive value} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$= \frac{\text{Number of death certificates that correctly attribute death to cancer } x}{\text{Total number of death certificates that attribute death to cancer } x}$$

$$Detection\ rate = \text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$= \frac{\text{Number of death certificates that correctly attribute death to cancer } x}{\text{True number of deaths due to cancer } x}$$

$$Correction\ factor = \frac{confirmation\ rate}{detection\ rate}$$

$$= \frac{\text{True number of deaths due to cancer } x}{\text{Total number of death certificates that attribute death to cancer } x}$$

Applying the Correction Factor to Adjust the Expected Number of Cancer Deaths

Since the SMR denominator represents the expected number of deaths due to cancer *x* based on general population cancer mortality rates derived exclusively from death certificate data, it follows that multiplying the SMR denominator by the correction factor will yield an approximation of the true number of expected deaths due to cancer *x*. For cancer types where the confirmation rate is superior to the detection rate, the correction factor is >1 and the corrected number of expected deaths will be higher than the number expected based solely on death

certificates. Conversely, for cancer types where the detection rate exceeds the confirmation rate, the correction factor is <1 and the corrected number of expected deaths will be lower than the number expected based only on death certificates.

Sensitivity Analyses

The results of SMR analyses performed with and without correction for discrepancies between death certificates and cancer registry records were generally very similar. Among men, the corrected and uncorrected SMR estimates for deaths due to cancers of the bronchus and lung, colon and rectum, and liver and intrahepatic bile ducts are shown below:

Cancer type	$SMR_{uncorrected}$	$SMR_{corrected}$
Bronchus and lung	2.37 (1.81-3.05)	2.39 (1.83-3.08)
Colon and rectum	2.17 (1.24-3.53)	2.37 (1.43-3.70)
Liver and intrahepatic bile ducts	4.20 (2.63-6.35)	4.35 (2.73-6.59)

Among men, 2 SMR estimates changed in statistical significance after correcting for discrepancies between death certificates and cancer registry records:

Cancer type	$SMR_{uncorrected}$	$SMR_{corrected}$
Larynx	4.04 (1.10-10.3)	3.13 (0.85-8.00)
Oral cavity and pharynx	2.23 (0.72-5.20)	2.37 (1.08-4.49)

The change in significance of the oral cavity and pharynx cancer estimate was due to an under-attribution of deaths from this cancer type on the death certificates of decedents in the homeless cohort when compared against the documented cancer site/type in their MCR case files. A similar pattern of misclassification has been observed in the general population, where death

certificates have a sensitivity of only 53% in detecting deaths due to oral cavity and pharynx cancer.[5] After upwardly revising both the SMR numerator and denominator to account for this pattern of under-identification, the SMR point estimate changed only slightly but achieved statistical significance because of an increase in the precision of the 95% CI.

Among women, the corrected and uncorrected SMR estimates for deaths due to cancers of the bronchus and lung, breast, and cervix uteri are shown below:

Cancer type	**$SMR_{uncorrected}$**	**$SMR_{corrected}$**
Bronchus and lung	2.33 (1.27-3.91)	2.31 (1.26-3.88)
Breast	1.15 (0.37-2.68)	1.07 (0.34-2.50)
Cervix uteri	7.52 (1.55-22.0)	6.01 (1.24-17.6)

A small number of deaths (1-4) attributed to liver and intrahepatic bile duct cancer on the death certificates of women appeared statistically greater than the number expected in uncorrected SMR analyses, but we subsequently reclassified all of these deaths to other cancer types based upon cross-examination with MCR records, yielding a corrected SMR of zero.

In view of the general similarity of the corrected and uncorrected results and the improved accuracy afforded by the correction methods, we report only the corrected results in the manuscript.

Appendix B References

1. CDC, National Center for Health Statistics. Underlying Cause of Death 1999-2011 on CDC WONDER Online Database, released 2014. Data are from the Multiple Cause of Death Files, 1999-2011, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. http://wonder.cdc.gov/ucd-icd10.html. Accessed September 23, 2014.
2. Gershman ST, Massachusetts Cancer Registry. Cancer Incidence and Mortality in Massachusetts 2004-2008: Statewide Report: Massachusetts Department of Public Health, Bureau of Health Information, Statistics, Research, and Evaluation; August 2011.
3. Hoel DG, Ron E, Carter R, Mabuchi K. Influence of death certificate errors on cancer mortality trends. *J Natl Cancer Inst.* 1993;85(13):1063-1068. http://dx.doi.org/10.1093/jnci/85.13.1063.
4. Ron E, Carter R, Jablon S, Mabuchi K. Agreement between death certificate and autopsy diagnoses among atomic bomb survivors. *Epidemiology.* 1994;5(1):48-56. http://dx.doi.org/10.1097/00001648-199401000-00009.
5. German RR, Fink AK, Heron M, et al. The accuracy of cancer mortality statistics based on death certificates in the United States. *Cancer Epidemiol.* 2010;35(2):126-131. http://dx.doi.org/10.1016/j.canep.2010.09.005.

Supplemental Table. ICD Codes Used for Identifying Incident Cancers (ICD-O-3) and Cancer Deaths (ICD-10).

Cancer site/type	ICD-O-3		ICD-10
	Topography	Morphology	
Brain and other nervous system	C70.0 - C72.9	All except 9590-9989	C70-C72
Breast	C50.0-C50.9	All except 9590-9989	C50
Bronchus and lung	C34.0-C34.9	All except 9590-9989	C34
Cervix uteri	C53.0-C53.9	All except 9590-9989	C53
Colon and rectum	C18.0-C18.9, C19.9, C20.9, C26.0	All except 9590-9989	C18-C20, C26.0
Corpus uteri and uterus NOS	C54.0-C54.9, C55.9	All except 9590-9989	C54-C55
Esophagus	C15.0-C15.9	All except 9590-9989	C15
Hodgkin lymphoma	C00.0-C80.9	9650-9667	C81
Kidney and renal pelvis	C64.9, C65.9	All except 9590-9989	C64-C65
Larynx	C32.0-C32.9	All except 9590-9989	C32
Leukemia	C00.0-C80.9	9733, 9742, 9800-9820, 9826, 9831-9948, 9963-9964	C90.1, C91-C95
	C42.0, C42.1, C42.4	9823, 9827	
Liver and intrahepatic bile ducts	C22.0, C22.1	All except 9590-9989	C22
Melanoma	C44.0-C44.9	All except 9590-9989	C43
Multiple myeloma	C00.0-C80.9	9731, 9732, 9734	C90.0, C90.2
Non-Hodgkin lymphoma	C00.0-C80.9	9590-9596, 9670-9729	C82-C85, C96.3
	All except C42.0, C42.1, C42.4	9823, 9827	
Oral cavity and pharynx	C00.0-C14.8	All except 9590-9989	C00-C14
Ovary	C56.9	All except 9590-9989	C56
Pancreas	C25.0-C25.9	All except 9590-9989	C25
Prostate	C61.9	All except 9590-9989	C61
Stomach	C16.0-C16.9	All except 9590-9989	C16
Testis	C62.0-C62.9	All except 9590-9989	C62
Thyroid	C73.9	All except 9590-9989	C73
Urinary bladder[a]	C67.0-C67.9	All except 9590-9989	C67

ICD-O-3, International Classification of Diseases for Oncology, 3rd Revision

[a] In keeping with Massachusetts Cancer Registry protocol, incident cases include *in situ* (behavior=2) and malignant (behavior=3) neoplasms of the urinary bladder. All other cancer types include only malignant (behavior=3) neoplasms.