ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO OPEN CANCER HORIZONS

ORIGINAL RESEARCH

Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit

A. Karlsson[1], A. Ellonen[2,3,4], H. Irjala[2,4,5], V. Väliaho[3,4], K. Mattila[3,4], L. Nissi[2,3], E. Kytö[2,4,5], S. Kurki[1,2], R. Ristamäki[3,4], P. Vihinen[4], T. Laitinen[6], A. Ålgars[2,3,4], S. Jyrkkiö[3,4], H. Minn[2,3,4] & E. Heervä[2,3,4*]

[1]Auria Biobank, University of Turku and Turku University Hospital, Turku; [2]University of Turku, Turku; [3]Department of Oncology, Turku University Hospital, Turku; [4]FICAN West Cancer Centre, Turku; [5]Department of Otorhinolaryngology—Head and Neck Surgery, Turku University Hospital, Turku; [6]Hospital Administration, Tampere University Hospital, Tampere, Finland

Available online 3 June 2021

Background: Persistent smoking after cancer diagnosis is associated with increased overall mortality (OM) and cancer mortality (CM). According to the 2020 Surgeon General's report, smoking cessation may reduce CM but supporting evidence is not wide. Use of deep learning-based modeling that enables universal natural language processing of medical narratives to acquire population-based real-life smoking data may help overcome the challenge. We assessed the effect of smoking status and within-1-year smoking cessation on CM by an in-house adapted freely available language processing algorithm.

Materials and methods: This cross-sectional real-world study included 29 823 patients diagnosed with cancer in 2009-2018 in Southwest Finland. The medical narrative, International Classification of Diseases-10th edition codes, histology, cancer treatment records, and death certificates were combined. Over 162 000 sentences describing tobacco smoking behavior were analyzed with ULMFiT and BERT algorithms.

Results: The language model classified the smoking status of 23 031 patients. Recent quitters had reduced CM [hazard ratio (HR) 0.80 (0.74-0.87)] and OM [HR 0.78 (0.72-0.84)] compared to persistent smokers. Compared to never smokers, persistent smokers had increased CM in head and neck, gastro-esophageal, pancreatic, lung, prostate, and breast cancer and Hodgkin's lymphoma, irrespective of age, comorbidities, performance status, or presence of metastatic disease. Increased CM was also observed in smokers with colorectal cancer, men with melanoma or bladder cancer, and lymphoid and myeloid leukemia, but no longer independently of the abovementioned covariates. Specificity and sensitivity were 96%/96%, 98%/68%, and 88%/99% for never, former, and current smokers, respectively, being essentially the same with both models.

Conclusions: Deep learning can be used to classify large amounts of smoking data from the medical narrative with good accuracy. The results highlight the detrimental effects of persistent smoking in oncologic patients and emphasize that smoking cessation should always be an essential element of patient counseling.

Key words: deep learning, artificial intelligence, tobacco use, cancer survival

INTRODUCTION

While smoking is a well-established risk factor for the development of many types of cancer, the 2020 report from the Surgeon General of the US Department of Health and Human Services emphasizes the need to assess the health effects of smoking cessation.[1] It has been estimated that two-thirds of patients who smoke continue to do so after cancer diagnosis.[2,3] This is alarming, since there is accumulating evidence that persistent smoking after cancer diagnosis impairs cancer-specific survival in multiple cancer types,[4] including small-cell and non-small-cell lung cancer (SCLC and NSCLC, respectively),[5,6] prostate,[7,8] head and neck,[9,10] colorectal,[11,12] and bladder cancer.[13] In breast cancer, conflicting results are reported, but large meta-analyses have demonstrated that persistent smoking is associated with poorer breast cancer-specific survival, especially in heavy smokers.[14,15]

Deep learning is an advanced subtype of artificial intelligence-based machine learning, which is making its way into clinical medicine via various diagnostic and predictive applications.[16,17] This includes universal language modeling of medical narratives, wherein real-life practice health data, such

*Correspondence to: Dr Eetu Heervä, Department of Oncology, Turku University Hospital, Hämeentie 11, Turku 20521, Finland. Tel: +35823130000; Fax: +35823131316

E-mail: eetu.heerva@utu.fi (E. Heervä).

as smoking status, are often presented in an unstructured format.[18] First reports of using language modeling to define an individual's smoking status appeared encouraging both from electronic health records (EHRs)[18,19] and from user-generated content on a smoking cessation support website.[20]

With the exception of lung and head and neck cancer, the effect of smoking cessation at the time of cancer diagnosis on cancer mortality (CM) remains inconclusive.[4,5,21,22] The aim of the current study was to evaluate the impact of smoking in cancer patients using deep learning, which allows us to efficiently study large patient populations. We applied this experimental setting on cancer patients diagnosed in the Turku University Hospital region, and subsequently analyzed (i) the effect of the individual's smoking status and (ii) the effect of recent smoking cessation on survival after cancer diagnosis.

MATERIALS AND METHODS

Ethics

The study was approved by the Institutional Review Board of Turku University Hospital, Turku, Finland (T132/2019). No informed consent was required according to the Finnish Act on Secondary Use of Health and Social Data. Based on the act, EHRs were reviewed and data were stored in a secured analysis environment protected by the hospital firewall.

Study population

Turku University Hospital is a tertiary referral Organization of European Cancer Institutes-designated cancer center covering the entire region with 480 000 inhabitants. The neighboring regions of Satakunta and Vaasa, with populations of 220 000 and 66 000, respectively, may refer patients to Turku for cancer care. Since 2004, Turku University Hospital uses a real-time updated EHR system. The EHRs were searched for all patients with at least one International Classification of Diseases-10th edition (ICD-10) C code, excluding non-melanoma skin cancer (C44). All patients whose first ICD-10 C code was recorded during 2009-2018 were included in the current study (Figure 1).

Definition of source data

The unique Finnish personal identification number was used to link patient-level data from different information systems, including ICD-10 codes, histology [International Classification of Diseases for Oncology-third edition (ICD-O-3) codes], Nordic operational codes [Nordic Medico-Statistical Committee (NOMESCO)], and chemotherapy [Anatomical Therapeutic Chemical (ATC) Classification codes] and radiotherapy records, to the medical narrative. For each patient, their medical narrative from 2004 to 2019 including on average 5600 words was extracted for language modeling. An extensive cross-linking of structured data was carried out to cross-validate all cancer diagnoses as accurately as possible. In case of discrepancy, such as a patient presenting with multiple ICD-10 codes, histology was used to establish the dominant diagnosis.

Date and cause of death were verified from Statistics Finland, an independent national registry. This verification also excluded non-Finnish citizens and ascertained complete follow-up. The number of incident cancer cases per region was obtained from the online statistics of the Cancer Registry of Finland (https://cancerregistry.fi/statistics/cancer-statistics/) to assess the population coverage.

EHRs were searched for body mass index (BMI), presence of select comorbidities, and Eastern Cooperative Oncology Group (ECOG) performance status. We defined synchronously metastatic cancer as presence of a metastasis code (C77-79) within 6 months of primary tumor diagnosis.

Language modeling for smoking status

For the current study, we built and used two language models, ULMFiT and Google BERT,[23,24] both of which use transfer learning. Firstly, a large amount of unlabeled data is leveraged to pre-train a model that performs well in the task of guessing masked words in phrases from the training data. The model learns a useful mathematical representation for the words and their connections. In this pre-training phase, the learned knowledge is transferred to a classifier supplemented with smaller amounts of manually labeled training data. The strength of this approach is that these models can be pre-trained with freely available Finnish text, and then manually fine-tuned with smaller clinical datasets, which in medicine are often hard and expensive to obtain.

Since no pre-trained models for Finnish were publicly available when we initiated the study, we pre-trained the ULMFiT model with the entire Finnish Wikipedia 2019 using the computer infrastructure at our hospital. For BERT, a pre-trained model for Finnish was later available.[24] The learned knowledge was then transferred to a training classifier, by randomly picking 5000 tobacco smoking-related sample phrases and sentences from the medical narrative archive of our hospital, using the Finnish word-stem 'tupak' equivalent to the English word-stem 'smok'.[19] These sample phrases were manually labeled into three classes (never, former, or current smoker). ULMFiT- and BERT-based classification models were then trained on this data to produce smoking phrase classifiers.

To define an individual's tobacco smoking status, a logic was added to both models when multiple classifications were present over time. Individuals labeled as both never and former smokers were classified as former smokers. In any other case, the average probability for each class over the patient's sentences was calculated for each smoking status classification, and the highest probability was used to determine the final status.

After language model classification, recent quitters were identified by extracting cessation year, where available, for ULMFiT-defined former smokers. Recent quitters were those who quit within 1 year before or any time after cancer diagnosis.[1]

Statistical analysis

The primary outcome measure was CM, defined as death due to any cancer. The secondary endpoint was overall mortality

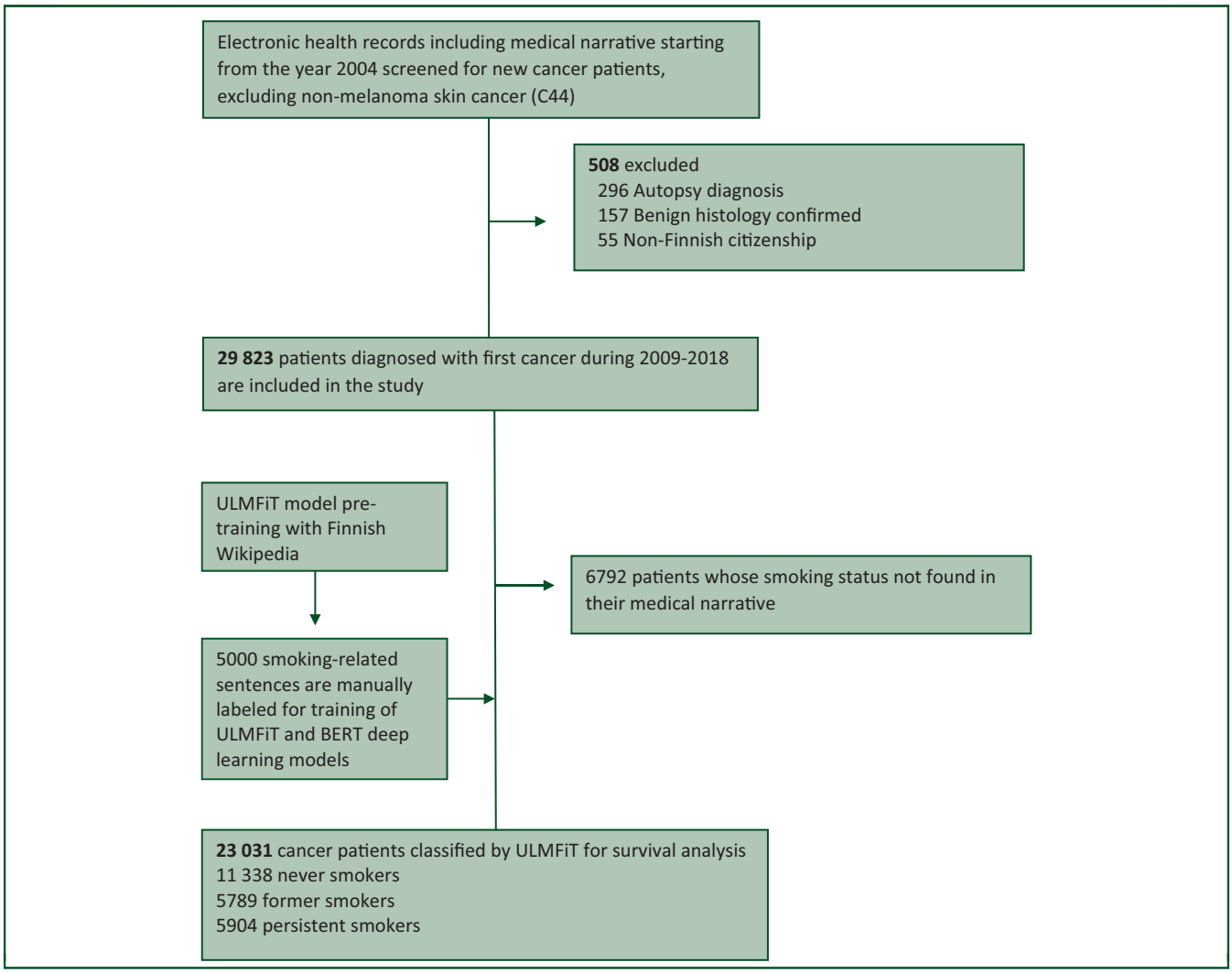

Figure 1. **Flowchart of the study design.**

(OM), defined as death due to any cause. All patients were censored at the time of death or at the end of follow-up in December 2019. Firstly, hazard ratios (HRs) were calculated with the Cox proportional hazards model with 95% confidence interval (CI_{95}), using sex as the categorical covariate with the enter method to control confounding. In case of significant sex interaction ($P < 0.05$), men and women were subsequently analyzed separately. Secondly, multivariate Cox regression was carried out, adjusting HRs for age >70 years, presence of synchronous metastasis (solid tumors only), ECOG 2 or more, presence of select comorbidities (Supplementary Table S1, available at https://doi.org/10.1016/j.esmoop.2021.100175), and BMI <20 kg/m^2. SPSS version 26 (IBM, Armonk, NY) was used. Sensitivity and specificity analyses were calculated with 2 × 2 contingency tables separately for never, former, and persistent smokers, excluding patients with missing smoking status,[19] using the Python 'sklearn' package.

RESULTS

The cohort consisted of 29 823 patients (14 717 females) with 35 394 cancers and a median (interquartile) age of 67 (58-76) years (Supplementary Table S1, available at https://doi.org/10.1016/j.esmoop.2021.100175). Median overall survival was not reached in female and 88.8 months in male cancer patients, with 5-year survival rates of 65% and 57%, respectively. At the end of follow-up, 12 244 patients had died with a minimum of 1 year of follow-up. Median follow-up was 5.2 years both in the entire series and in those alive at the end of the study. Good population coverage was observed as compared to national data from the Cancer Registry of Finland, being lowest in lung (81%), renal (92%), and colorectal cancer (93%, Supplementary Table S2, available at https://doi.org/10.1016/j.esmoop.2021.100175). When medical narratives were analyzed, we found on average 7 hits (range 0-116) per patient for sentences that describe patient's smoking behavior. The ULMFiT language model classified a total of 11 338 never, 5789 former, and 5904 persistent smokers. For 6792 patients, especially in ovarian and endometrial cancers, we found no mention of smoking (Supplementary Table S3, available at https://doi.org/10.1016/j.esmoop.2021.100175).

Language model performance

The overall precision of the ULMFiT model was 87.4% compared to 88.2% for Google BERT based on the blinded

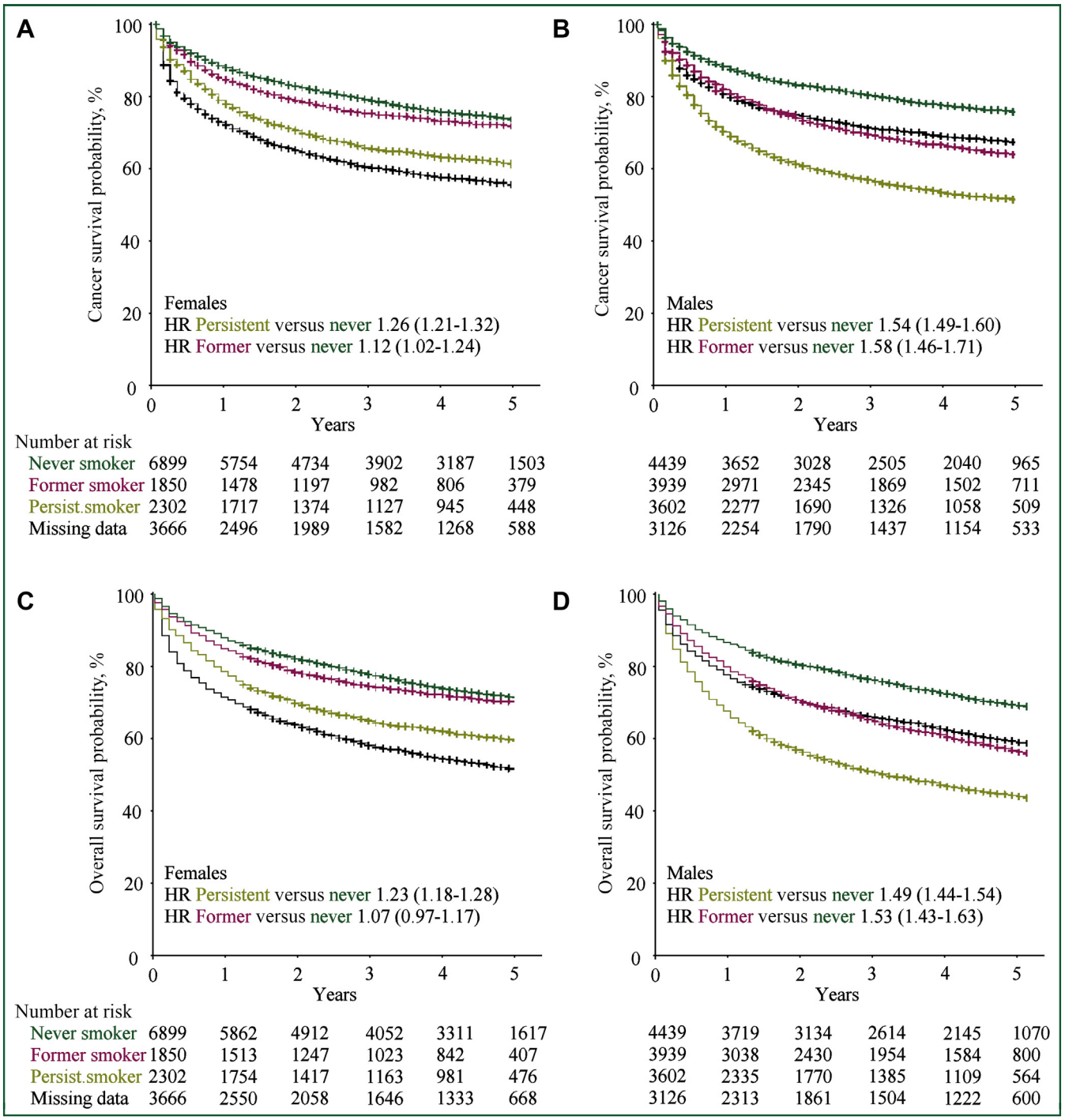

Figure 2. Kaplan–Meier graphs of 5-year cancer (A and B) and overall mortality (C and D) according to sex.
Hazard ratios (HRs) with 95% confidence intervals (CI_{95}) are calculated with Cox regression analysis. Persist., persistent.

manual classification of smoking status of 1014 patients (Supplementary Table S4, available at https://doi.org/10.1016/j.esmoop.2021.100175). Specificity and sensitivity with the ULMFiT model were 96%/96%, 98%/68%, and 88%/99% for never, former, and current smokers, respectively. With the BERT model, the respective specificity and sensitivity were 96%/96%, 96%/73%, and 90%/97%. The discrepant cases were elderly people who had tried smoking in their youth, smokers attempting to quit, or those where only the most recent EHR entry suggested successful smoking cessation.

Smoking impairs cancer and overall mortality

In male patients with cancer, as compared to never smokers, both persistent [HR 1.54 (1.49-1.60)] and former smoking [HR 1.58 (1.46-1.71)] increased CM (Figure 2). Similar results, but smaller in size, were observed in female

	n (never smoker)	*n* (persistent smoker)	HR
Head and neck (C00-14, 30-32)	333	458	**CM 1.46 (1.27-1.68)**
			OM 1.41 (1.25-1.59)
Gastro-esophageal (C15-16)	234	241	**CM 1.22 (1.10-1.36)**
			OM 1.24 (1.12-1.37)
Colorectal (C18-20)	1090	408	**CM 1.13 (1.03-1.25)**
			OM 1.14 (1.04-1.24)
Pancreatic (C23-25)	411	272	CM 1.05 (0.97-1.14)
			OM 1.04 (0.96-1.13)
NSCLC (C34)	227	907	**CM 1.17 (1.06-1.28)**
			OM 1.17 (1.07-1.28)
SCLC (C34)	31	246	**CM 1.64 (1.25-2.15)**
			OM 1.63(1.26-2.12)
Cutaneous melanoma (C43), females	327	63	CM 1.25 (0.88-1.78)
			OM 1.24 (0.91-1.68)
Cutaneous melanoma (C43), males	273	89	**CM 1.40 (1.13-1.74)**
			OM 1.30 (1.07-1.58)
Breast (C50)	2938	765	CM 0.97 (0.85-1.11)
			OM 0.95 (0.86-1.05)
Prostate (C61)	1712	725	**CM 1.31 (1.17-1.48)**
			OM 1.23 (1.13-1.34)
Renal (C64)	288	142	CM 1.07 (0.89-1.28)
			OM 1.10 (0.94-1.29)
Bladder (C65-68), females	104	49	CM 0.85 (0.59-1.25)
			OM 0.97 (0.72-1.30)
Bladder (C65-68), males	184	266	**CM 1.21 (1.01-1.49)**
			OM 1.21 (1.03-1.41)
Brain (C71)	263	103	CM 0.93 (0.80-1.07)
			OM 0.92 (0.79-1.06)
Hodgkin's (C81)	104	45	**CM 1.70 (1.05-2.77)**
			OM 1.35 (0.88-2.07)
Non-Hodgkin's (C82-85)	533	185	CM 1.05 (0.90-1.23)
			OM 1.09 (0.95-1.25)
Myeloma (C90)	209	72	CM 0.95 (0.75-1.20)
			OM 1.07 (0.89-1.30)
Lymphoid leukemia (C91)	177	61	**CM 1.41 (1.02-1.95)**
			OM 1.41 (1.09-1.83)
Myeloid leukemia (C92)	141	57	**CM 1.27 (1.02-1.57)**
			OM 1.27 (1.03-1.58)

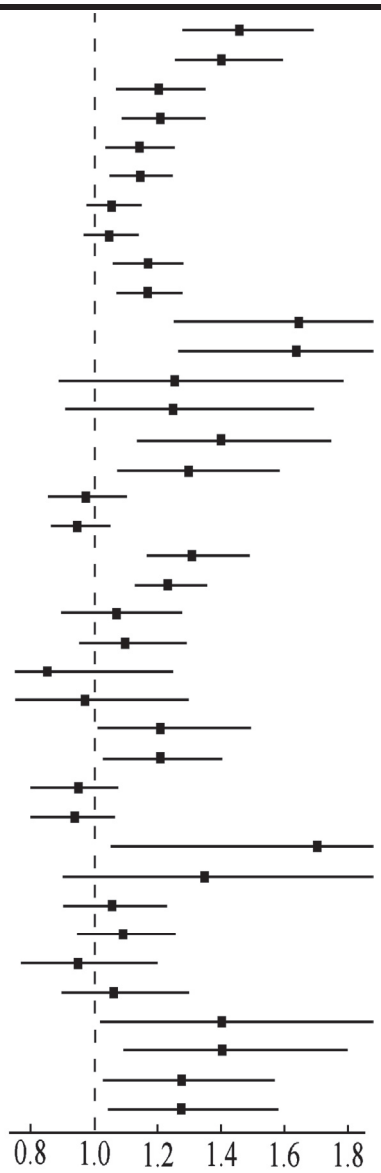

Figure 3. Persistent smoking and survival.
Cox regression analysis of cancer (CM) and overall mortality (OM) with sex as covariate. Where significant (bold), men and women are analyzed separately.
HR, hazard ratio; NSCLC, non-small-cell lung cancer; SCLC, small-cell lung cancer.

persistent [HR 1.26 (1.21-1.32)] and former smokers [HR 1.12 (1.02-1.24), Figure 2]. Results with OM appeared similar but were of lesser magnitude (Figure 2).

Specifically, persistent smokers had increased CM in head and neck cancer, gastro-esophageal cancer, colorectal cancer, NSCLC, SCLC, cutaneous melanoma (men only), prostate cancer, bladder cancer (men only), Hodgkin's lymphoma, and both lymphoid and myeloid leukemia (Figure 3). In former smokers, a trend favoring higher CM was observed in the abovementioned cancers, being significant in NSCLC, SCLC, bladder and prostate cancer, and lymphoid leukemia (Supplementary Figure S1, available at https://doi.org/10.1016/j.esmoop.2021.100175). Never smokers were predominantly female, and smokers were on average 3 years younger and had more often poor performance status as compared to others (Supplementary Table S1, available at https://doi.org/10.1016/j.esmoop.2021.100175).

HRs of OM rarely exceeded HRs of CM, with the exceptions of myeloma, females with bladder cancer, and former smokers with cutaneous melanoma and renal cancer (Figure 3, Supplementary Figure S1, available at https://doi.org/10.1016/j.esmoop.2021.100175).

After CM HRs for persistent and former smokers were adjusted for age, presence of synchronous metastasis, ECOG, comorbidities, and BMI, persistent smoking was an independent risk factor for cancer-related death in head

Table 1. Adjusted HRs for cancer mortality

Cancer type	*n* (never smoker)	*n* (former smoker)	*n* (persistent smoker)	HR (CI_{95}) persistent versus never	HR (CI_{95}) former versus never
Head and neck	333	287	458	**1.25 (1.04-1.49)**	1.07 (0.71-1.62)
Gastro-esophageal	235	203	242	**1.17 (1.00-1.37)**	1.15 (0.83-1.59)
Colorectal	1090	545	408	1.04 (0.92-1.17)	0.87 (0.69-1.08)
Pancreatic	411	204	272	**1.16 (1.03-1.29)**	1.06 (0.84-1.33)
NSCLC	227	486	907	**1.34 (1.17-1.52)**	**1.50 (1.13-1.98)**
SCLC	31	65	246	**1.62 (1.17-2.24)**	**2.16 (1.04-4.48)**
Cutaneous melanoma	600	203	152	1.19 (0.94-1.50)	0.90 (0.59-1.37)
Breast	2938	862	765	**1.20 (1.02-1.40)**	0.98 (0.72-1.33)
Prostate	1712	1288	725	**1.22 (1.03-1.45)**	1.07 (0.81-1.41)
Renal	288	131	142	1.15 (0.88-1.49)	0.81 (0.47-1.38)
Bladder	288	283	315	0.99 (0.77-1.29)	1.32 (0.79-2.21)
Brain	263	109	103	0.78 (0.63-1.00)	0.95 (0.67-1.33)
Hodgkin's	104	38	45	**2.45 (1.09-5.93)**	3.16 (0.78-12.73)
Non-Hodgkin's	533	273	185	1.08 (0.90-1.31)	**1.38 (1.01-1.89)**
Myeloma	209	98	72	1.17 (0.70-1.96)	1.24 (0.66-2.35)
Lymphoid leukemia	177	67	61	1.11 (0.56-2.20)	2.37 (0.75-7.49)
Myeloid leukemia	144	61	57	1.40 (0.95-2.05)	0.39 (0.10-1.48)

Cox multivariate analysis includes presence of synchronous metastasis (solid tumors only), age >70 years, sex, ECOG 2 or more, body mass index <20 kg/m^2, and presence of select comorbidities (diabetes, coronary heart disease, cerebrovascular disease, chronic obstructive pulmonary disease, and liver failure). Significant results are given in bold.
CI_{95}, 95% confidence interval; HR, hazard ratio; NSCLC, non-small-cell lung cancer; SCLC, small-cell lung cancer.

Table 2. CM and OM of recent quitters compared to persistent smokers

Cancer type	*n* (recent quitter)	*n* (persistent smoker)	CM, HR (CI_{95})	OM, HR (CI_{95})
All	524	5863	**0.80 (0.74-0.87)**	**0.78 (0.72-0.84)**
Head and neck (C00-14, 30-32)	65	441	**0.65 (0.50-0.85)**	**0.60 (0.47-0.76)**
Gastro-esophageal (C15-16)	20	238	**0.75 (0.56-0.99)**	**0.70 (0.53-0.93)**
Colorectal (C18-20)	36	402	0.76 (0.55-1.07)	**0.71 (0.52-0.97)**
Pancreatic (C23-25)	16	272	0.90 (0.68-1.21)	0.96 (0.74-1.26)
NSCLC (C34)	67	899	**0.84 (0.73-0.97)**	**0.83 (0.72-0.96)**
SCLC (C34)	12	245	0.91 (0.67-1.23)	0.89 (0.66-1.20)
Cutaneous melanoma (C43)	14	152	0.65 (0.32-1.31)	0.85 (0.51-1.41)
Breast (C50)	82	764	0.91 (0.60-1.39)	0.88 (0.63-1.23)
Prostate (C61)	68	725	0.93 (0.68-1.27)	0.85 (0.66-1.10)
Renal (C64)	11	142	0.94 (0.56-1.56)	0.86 (0.55-1.37)
Bladder (C65-68)	20	315	1.15 (0.80-1.65)	0.93 (0.65-1.32)
Brain (C71)	9	103	0.73 (0.41-1.30)	0.70 (0.40-1.25)
Hodgkin's (C81)	13	43	0.55 (0.19-1.54)	0.54 (0.19-1.55)
Non-Hodgkin's (C82-85)	17	185	0.81 (0.49-1.34)	0.80 (0.51-1.26)
Leukemias combined (C91-92)	6	118	0.87 (0.43-1.77)	0.97 (0.54-1.74)

Patients who were manually verified as recent quitters are moved into recent quitter cohort. Significant results are given in bold.
CI_{95}, 95% confidence interval; CM, cancer mortality; HR, hazard ratio; NSCLC, non-small-cell lung cancer; OM, overall mortality; SCLC, small-cell lung cancer.

and neck, gastro-esophageal, pancreatic, NSCLC, SCLC, breast, and prostate cancer and in Hodgkin's lymphoma. Thus, smoking emerged as a significant covariate in pancreatic and breast cancer, while significance was no longer observed in cutaneous melanoma and colorectal cancer (Table 1).

Recent quitting reduces cancer mortality

Smoking cessation year could be extracted for 2803 (48%) former smokers. An additional 41 recent quitters were identified during the manual validation process; thus, a total of 524 patients met the criteria for recent quitter.[1] Compared to those who continue smoking, recent quitting reduced both CM [HR 0.80 (0.74-0.87), Table 2] and OM [HR 0.78 (0.72-0.84), Table 2]. A tendency for lower CM and OM was observed almost in every cancer type except for bladder cancer, where only reduced OM after smoking cessation was observed. Interestingly, also in the sub-group of 2282 patients with synchronous metastatic disease, reduced OM was observed in 53 recent quitters as compared to 759 persistent smokers with a HR of 0.85 (0.72-0.98).

DISCUSSION

The recent Surgeon General's report emphasized the need to assess the health effects of smoking cessation at the time of cancer diagnosis.[1] Motivated by this report, we applied deep learning-based language modeling to analyze the medical narratives of a large cross-sectional cohort of cancer patients. With good population-based coverage, we confirmed the detrimental effect of persistent smoking on CM and OM in many types of cancer (head and neck, colorectal, NSCLC, SCLC, melanoma, breast, prostate, bladder, and myeloid leukemia). Interestingly, persistent smoking increased CM in gastro-esophageal and pancreatic

cancer, lymphoid leukemia, and Hodgkin's lymphoma, to our knowledge not previously extensively reported in European populations. Furthermore, a major finding was that patients who quit smoking not earlier than 1 year before diagnosis showed significantly improved cancer and overall survival, including those with synchronously metastatic disease.

One major strength of the current study is that we were able to link extensive structured clinical data to smoking data, and showed that smoking was an independent modifiable risk factor for CM in multiple cancer types. While the detrimental effects of persistent smoking have been shown before,[4,21] our results suggest that smoking cessation is beneficial irrespective of patient's age, comorbidities, or metastatic disease. We also analyzed OM along with CM and observed that OM HRs rarely exceeded CM HRs, suggesting that cancer patients who continue smoking will probably die due to cancer and not from other tobacco-related diseases.

Artificial intelligence and deep learning applications are making their way into clinical practice, but their efficacy is not yet proven in prospective trial settings.[16] The current study utilized more advanced deep learning compared to previous reports[18-20] in a retrospective proof-of-principle setting, where we successfully extracted a large amount of smoking data in a matter of days. Two language models were tested with good specificity (88%-98%), comparable to the previous results in English language.[19] The BERT model included a pre-trained model for Finnish,[24] but with seemingly only slightly improved specificity in persistent smokers compared to ULMFiT. Discrepancies were most commonly due to unsuccessful attempts to quit, and the individual's smoking status was ultimately based on probability logic instead of on the language model itself. Misclassifications also had human reasons behind them; elderly people who tried smoking in their youth are former smokers by Definition, while human classification may overlook sporadic smoking. Furthermore, the amount of missing data was almost always true missing data, which may reflect healthcare professionals' current practices and attitudes toward reporting smoking behavior of their patients.[25]

In the current study, 524 (8%) smokers at the time of cancer diagnosis managed to quit smoking and remain abstinent. This 8% is probably an underestimate, since the cessation year was found only in 48% of former smokers, but may reflect the hardships of smoking cessation in a real-world population. For comparison, abstinence rates of 36%-43% have been reported in large surveys and smoking cessation programs.[2,3,22] Regardless of missing cessation years, our study showed a reduced CM and OM in recent quitters compared to persistent smokers. We observed that former smokers were more often ECOG 0-1 compared to smokers, without explanatory sex or age differences.

One novel observation in our study is that smoking was an independent risk factor for Hodgkin's lymphoma mortality. However, the current study was not specifically designed to address predictive factors in lymphomas, and survival may be affected by a variety of unavailable covariates, including staging of hematological malignancies. The same holds also true for lymphoid leukemia, while in myeloid leukemia the detrimental effect of smoking has been reported earlier.[26] While smoking is a well-known adverse factor in esophageal cancer,[1] in pancreatic and gastric cancer smoking data are scarce or derived from Asian populations.[27,28] Results on cutaneous melanoma should be observed with caution, since detrimental effects of smoking were observed in men only and not significant once adjusted for other covariates, but are in line with previous reports.[4,21]

The limitations of our study include its retrospective single-center nature and that we used only Finnish texts. However, ULMFiT and BERT are designed as universal language models, and the same methodology can be applied in any language provided enough data are available.[19,23,24] We acknowledge that our results should be reproduced in different languages before wider implementation. A possible limitation of our study design is that language models such as ours are trained on ordinary, everyday language from Wikipedia pages and discussion forums. Medical text is typically laden with technical terms that do not appear in everyday discussions, which could inhibit the transferability of what the model has learned in the pre-training phase.

Conclusions

To our knowledge, this is the largest comprehensive deep learning-based registry of smoking behaviors in cancer patients. We applied this experimental setting to link the individual's smoking status to cancer outcome and demonstrated the detrimental effects of persistent smoking on CM across multiple types of cancer independent of age, presence of metastatic disease, BMI, performance status, or comorbidities. The current study suggests that cancer survival may be improved through smoking cessation, and it is never too late to quit. These results encourage a broader international validation of language models in clinical practice and emphasize that smoking counseling should always be an integral part of cancer care.

ACKNOWLEDGEMENTS

The authors thank everyone currently working or formerly worked at Auria Clinical Informatics for their data gathering, curation, and storage. Antti Sykkö and Heidi Kurri from Statistics Finland are thanked for the cause of death records and Adelaide Lönnberg for revising the English language.

FUNDING

This work was supported by the Cancer Society of Finland and State Competitive Research Funds (no grant number). The funders had no role in the study design, interpretation, or writing of the text.

DISCLOSURE

The authors have declared no conflicts of interest.

DATA SHARING

Patient-level data, even if deidentified or summarized, cannot be transferred without an internal permission procedure according to the Finnish legislation on the secondary use of health data (see 552/2019, www.finlex.fi). Data may be requested from www.auria.fi. The pre-trained ULMFiT language model is available at https://github.com/AnttiKarlsson/finnish_ulmfit. The pre-trained Finnish BERT model is available at http://turkunlp.org/FinBERT/.

REFERENCES

1. U.S. Department of Health and Human Services. *Smoking Cessation. A Report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2020.
2. Tseng TS, Lin HY, Moody-Thomas S, Martin M, Chen T. Who tended to continue smoking after cancer diagnosis: the national health and nutrition examination survey 1999—2008. *BMC Public Health*. 2012;12:784.
3. Cinciripini PM, Karam-Hage M, Kypriotakis G, et al. Association of a comprehensive smoking cessation program with smoking abstinence among patients with cancer. *JAMA Netw Open*. 2019;2:e1912251.
4. Wang Y, Tao H, Paxton RJ, et al. Post-diagnosis smoking and risk of cardiovascular, cancer, and all-cause mortality in survivors of 10 adult cancers: a prospective cohort study. *Am J Cancer Res*. 2019;9:2493-2514.
5. Parsons A, Daley A, Begh R, Aveyard P. Influence of smoking cessation after diagnosis of early stage lung cancer on prognosis: systematic review of observational studies with meta-analysis. *Br Med J*. 2010;340:b5569.
6. Gemine RE, Ghosal R, Collier G, et al. Longitudinal study to assess impact of smoking at diagnosis and quitting on 1-year survival for people with non-small cell lung cancer. *Lung Cancer*. 2019;129:1-7.
7. Islami F, Moreira DM, Boffetta P, Freedland SJ. A systematic review and meta-analysis of tobacco use and prostate cancer mortality and incidence in prospective cohort studies. *Eur Urol*. 2014;66:1054-1064.
8. Darcey E, Boyle T. Tobacco smoking and survival after a prostate cancer diagnosis: a systematic review and meta-analysis. *Cancer Treat Rev*. 2018;70:30-40.
9. Sharp L, McDevitt J, Carsin AE, Brown C, Comber H. Smoking at diagnosis is an independent prognostic factor for cancer-specific survival in head and neck cancer: findings from a large, population-based study. *Cancer Epidemiol Biomarkers Prev*. 2014;23:2579-2590.
10. Lassen P, Lacas B, Pignon JP, et al. Prognostic impact of HPV-associated p16-expression and smoking status on outcomes following radiotherapy for oropharyngeal cancer: the MARCH-HPV project. *Radiother Oncol*. 2018;126:107-115.
11. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal cancer: a meta-analysis. *J Am Med Assoc*. 2008;300:2765-2778.
12. Walter V, Jansen L, Hoffmeister M, Brenner H. Smoking and survival of colorectal cancer patients: systematic review and meta-analysis. *Ann Oncol*. 2014;25:1517-1525.
13. Cumberbatch MG, Rota M, Catto JW, La Vecchia C. The role of tobacco smoke in bladder and kidney carcinogenesis: a comparison of exposures and meta-analysis of incidence and mortality risks. *Eur Urol*. 2016;70:458-466.
14. Duan W, Li S, Meng X, Sun Y, Jia C. Smoking and survival of breast cancer patients: a meta-analysis of cohort studies. *Breast*. 2017;33:117-124.
15. Sollie M, Bille C. Smoking and mortality in women diagnosed with breast cancer—a systematic review with meta-analysis based on 400,944 breast cancer cases. *Gland Surg*. 2017;6:385-393.
16. Panagiotou OA, Högg LH, Hricak H, et al. Clinical application of computational methods in precision oncology: a review. *JAMA Oncol*. 2020;6:1282-1286.
17. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. 2019;20:e262-e273.
18. Lin FP, Pokorny A, Teng C, Epstein RJ. TEPAPA: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Sci Rep*. 2017;7:6918.
19. Palmer EL, Hassanpour S, Higgins J, Doherty JA, Onega T. Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC Med Inform Decis Mak*. 2019;19:141.
20. Wang X, Zhao K, Cha S, et al. Mining user-generated content in an online smoking cessation community to identify smoking status: a machine learning approach. *Decis Support Syst*. 2019;116:26-34.
21. Warren GW, Kasza KA, Reid ME, Cummings KM, Marshall JR. Smoking at diagnosis and survival in cancer patients. *Int J Cancer*. 2013;132:401-410.
22. Day AT, Dahlstrom KR, Lee R, Karam-Hage M, Sturgis EM. Impact of a tobacco treatment program on abstinence and survival rates among current smokers with head and neck squamous cell carcinoma. *Head Neck*. 2020;42:2440-2452.
23. Eisenschloss JM, Ruder S, Czapla P, Kardas M, Gugger S, Howard J. MultiFIT: efficient multi-lingual language model fine-tuning. Available at https://arxiv.org/abs/1909.04761. Accessed February 18, 2021.
24. Virtanen A, Kanerva J, Ilo R, et al. Multilingual is not enough: BERT for Finnish. Available at https://arxiv.org/pdf/1912.07076. Accessed February 18, 2021.
25. Derksen JWG, Warren GW, Jordan K, et al. European practice patterns and barriers to smoking cessation after a cancer diagnosis in the setting of curative versus palliative cancer treatment. *Eur J Cancer*. 2020;138:99-108.
26. Chelghoum Y, Danaïla C, Belhabri A, et al. Influence of cigarette smoking on the presentation and course of acute myeloid leukemia. *Ann Oncol*. 2002;13:1621-1627.
27. Yuan C, Morales-Oyarvide V, Babic A, et al. Cigarette smoking and pancreatic cancer survival. *J Clin Oncol*. 2017;35:1822-1828.
28. Minami Y, Kanemura S, Oikawa T, et al. Associations of cigarette smoking and alcohol drinking with stomach cancer survival: a prospective patient cohort study in Japan. *Int J Cancer*. 2018;143:1072-1085.