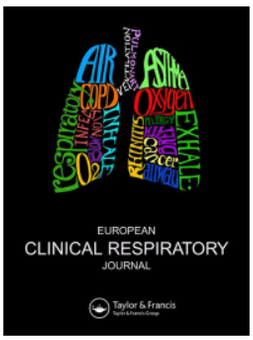

European Clinical Respiratory Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/zecr20

Documentation of the patient's smoking status in common chronic diseases – analysis of medical narrative reports using the ULMFiT based text classification

Eveliina Hirvonen, Antti Karlsson, Tarja Saaresranta & Tarja Laitinen

To cite this article: Eveliina Hirvonen, Antti Karlsson, Tarja Saaresranta & Tarja Laitinen (2021) Documentation of the patient's smoking status in common chronic diseases – analysis of medical narrative reports using the ULMFiT based text classification, European Clinical Respiratory Journal, 8:1, 2004664, DOI: 10.1080/20018525.2021.2004664

To link to this article: https://doi.org/10.1080/20018525.2021.2004664

Published online: 23 Nov 2021.

EUROPEAN CLINICAL RESPIRATORY JOURNAL
2021, VOL. 8, 2004664
https://doi.org/10.1080/20018525.2021.2004664

Taylor & Francis
Taylor & Francis Group

Documentation of the patient's smoking status in common chronic diseases – analysis of medical narrative reports using the ULMFiT based text classification

Eveliina Hirvonen [a,b], Antti Karlsson[c], Tarja Saaresranta[a,b] and Tarja Laitinen[a,d]

[a]Division of Medicine, Department of Pulmonary Diseases, Turku University Hospital, Turku, Finland; [b]Department of Pulmonary Diseases and Clinical Allergology, University of Turku Turku Finland; [c]Auria Biobank, University of Turku and Turku University Hospital, Turku, Finland; [d]Administration Centre, Tampere University Hospital, Tampere, Finland

ABSTRACT

Introduction: Smoking cessation is essential part of a successful treatment in many chronic diseases. Our aim was to analyse how actively clinicians discuss and document patients' smoking status into electronic health records (EHR) and deliver smoking cessation assistance.

Methods: We analysed the results using a combination of rule and deep learning-based algorithms. Narrative reports of all adult patients, whose treatment started between years 2010 and 2016 for one of seven common chronic diseases, were followed for two years. Smoking related sentences were first extracted with a rule-based algorithm. Subsequently, pre-trained ULMFiT-based algorithm classified each patient's smoking status as a current smoker, ex-smoker, or never smoker. A rule-based algorithm was then again used to analyse the physician-patient discussions on smoking cessation among current smokers.

Results: A total of 35,650 patients were studied. Of all patients, 60% were found to have a smoking status in EHR and the documentation improved over time. Smoking status was documented more actively among COPD (86%) and sleep apnoea (83%) patients compared to patients with asthma, type 1&2 diabetes, cerebral infarction and ischemic heart disease (range 44-61%). Of the current smokers (N=7,105), 49% had discussed smoking cessation with their physician. The performance of ULMFiT-based classifier was good with F-scores 79-92.

Conclusion: Ee found that smoking status was documented in 60% of patients with chronic disease and that the clinician had discussed smoking cessation in 49% of patients who were current smokers. ULMFiT-based classifier showed good/excellent performance and allowed us to efficiently study a large number of patients' medical narratives.

ARTICLE HISTORY
Received 23 May 2021
Accepted 6 November 2021

KEYWORDS
Smoking; smoking cessation; smoking intervention; electronic health records; medical narrative; natural language processing; machine learning; deep learning; artificial intelligence; language modelling; transfer learning; ULMFiT

Introduction

Smoking continues to be the leading preventable cause of death and illness, causing 8 million premature deaths each year [1]. In Finland, 14% of adults smoke daily [2]. Smoking is a clear risk factor for initiation and progression of several diseases and often affects long-term treatment outcomes [3,4]. Therefore, all clinical guidelines recommend that the risks of smoking should be discussed with patients. It is also important that the conversations are well documented. For health professionals, it is crucial to include this piece of information in their routine care in order to reliably assess the risks and efficacy of the treatment and to provide smoking cessation assistance. However, physicians often underuse these opportunities to deliver cessation intervention to smokers [5,6]. Already, a short discussion with a patient has been shown to increase the likelihood of quitting [7,8] with combined behavioral support and pharmacotherapy being the most effective [9].

In Finnish electronic health records (EHR), smoking is usually documented as free text. This makes it challenging for secondary use of EHRs when, for example, the effectiveness of given treatments is evaluated. The development of natural language processing (NLP) technologies has improved these processes [10] but, due to the complex nature of clinical phrases and expressions, the applications have proven challenging. Even more obstacles are encountered when these applications have been transferred to other languages beyond English. In recent years, deep learning-based approaches have brought new solutions for NLP tasks but the algorithms still need large training sets to be valid. One solution is to utilize transfer learning [11]. In 2018, Howard and Ruder developed the Universal

CONTACT Eveliina Hirvonen rievhi@utu.fi Division of Medicine, Department of Pulmonary Diseases, Turku University Hospital, P.O.BOX 52 (Hämeentie 11), 20521 Turku, Finland

Language Model Fine-tuning (ULMFiT) model a transfer learning method that can be used for various NLP tasks, including text classification [12]. The idea is to first pre-train the language model to learn the structure and general features of the study language. This can be done without any labels, for example, with a massive text data obtained from the internet. Then, the model is fine-tuned with a domain-specific language. Finally, a classification model is built on top of the fine-tuned language model using a labelled dataset. Thus, knowledge is transferred from the pre-training phase into the classifier that results in better classifiers than the ones trained on the labelled, usually much smaller dataset alone [10,11]. Besides ULMFiT, several similar models have been published, for example, by Google [13], Facebook [14] and OpenAI [15].

The aim of our study was to examine the documentation of the discussions between the clinician and the patient about smoking and in case of current smokers, the encouragement of smoking cessation. We wanted to also know whether these practices have improved at the Turku University Hospital over a 9-year period. In total, we studied seven common chronic diseases in which smoking cessation is an essential part of the treatment. We also investigated the performance of a ULMFiT-based algorithm in classifying the patients into never, ex – or current smokers.

Methods

Study cohort and data source

The narrative reports of the patients who were >18 years of age and diagnosed between the years 2010 and 2016 with asthma (ICD10 codes J45-46), chronic obstructive pulmonary disease (COPD, J44), type 1 diabetes (E10), type 2 diabetes (E11), sleep apnoea (G47), ischemic heart diseases (IHD I20-25), or cerebral infarction (I63) at the Turku University Hospital were included. Patients were either diagnosed for the first time with the disease or they were referred to secondary health care for treatment optimization. The patients' medical narratives were then followed for two years starting from the first visit. Majority of the narrative reports, which are part of EHR, were made by physician, but small number of narratives were made by other healthcare professionals. If a patient had two or more of above-mentioned diagnoses, the patient was only included in the group defined by the diagnosis that appeared first, i.e. a patient was first treated for this disease at the study hospital.

Identification of the patients' smoking status

All Finnish sentences specific to smoking and tobacco were extracted from the medical narratives using the rule-based algorithm 1. These sentences were then analysed using the ULMFiT-based algorithm that classified smoking-related phrases into three classes: current smoker, ex-smoker, or never smoker. The algorithm was pre-trained using the Finnish Wikipedia 2019 and then finetuned using the Finnish narrative reports of 5,000 cancer patients from the same hospital. The narrative reports were manually annotated into the same three classes [16]. In addition, a total of 40 random patients in each disease group studied were classified in a similar fashion by a physician (EH) in order to validate the performance of the algorithm in these particular disease groups. If the patient's smoking status changed over time, the most frequently appearing status was included in the study.

Identification of the patients who were encouraged to quit smoking

In a similar manner, ad hoc rules for the Finnish terms related to an encouragement to quit smoking were used to extract corresponding sentences. The algorithm was manually validated using a random sample of 50 + 50 patients classified as current smokers and either being or not being encouraged to quit smoking. Based on patients' medical records, we also evaluated the number of visits to the nurse-managed smoking cessation program at the Turku University Hospital.

This retrospective, registry-based study approach was approved by the administration of the Turku University Hospital (number T316/2019). The data was stored and analyzed in a secured IT environment owned by the Turku University Hospital. Only the study team had access to the data through 2-factor authentication.

Statistical analyses

Statistical analyses were performed with Excel for Mac 2018. The figure was made using Excel for Mac 2018. Continuous variables were presented as means and standard deviation (SD) for normally distributed variables or median and interquartile range (IQR) for non-normally distributed variables. Categorical variables were presented as frequencies and proportions. Statistical comparison between the groups was carried out using a chi-squared test. Statistical significance was considered as a p-value <0.05.

The performance of the algorithms was assessed based on accuracy, precision, recall and F1-score. We built a 2 × 2 confusion matrix with the following labels: true positive (TP), true negative (TN), false-positive (FP) and false-negative (FN). We compared the algorithms' results to the physician's classification (true values). Accuracy describes the proportion of true values ((TP+TN)/ n), precision defines the accuracy of positive values (TP/(TP+FP)), recall is the fraction of correctly predicted true values of all true values (TP/(TP+FN) and F1-score combines precision and recall to a single value (2/(1/precision+1/recall)). The performance of the ULMFiT algorithm was evaluated separately for current, ex – and never smokers.

Results

Characteristics of the study cohort

Based on our approach we identified a total of 4,549 adult asthma, 2,111 COPD, 5,931 sleep apnoea, 632 type 1 diabetes, 8,281 type 2 diabetes, 9,200 IHD, and 4,946 cerebral infarction patients (Table 1). The median length of their 2-year medical narrative after the given diagnosis varied from 12 to 28 events consisting mainly of inpatient and outpatient visits.

Smoking status in different disease groups and over time

Within the total cohort, 40.1% of all patient's medical documentation failed to reveal any smoking-related phrases during their two-year medical narrative. When comparing the documentation since the years 2010–2012 to 2016–2018, some improvement was observed especially in patients with cerebral infarction (344/695 and 453/674, +18%, p = 0.001, respectively) (Figure 1). A patient's smoking status was documented significantly more frequently in COPD (86%) and sleep apnoea patients (83%) compared to other groups (84% vs 53%, p < 0.001). In general, health professionals had documented discussions on smoking with the ex – and current smokers more often than with the never smokers during a two-year follow-up.

The proportion of never smokers varied between disease groups from 3% to 53% (Table 1). Of COPD patients, 70% were classified as the current smokers in comparison to 25–36% in the other disease groups (Table 1). Overall, the proportions of patients classified as current smokers decreased in all disease groups over the 9-year observation period. We compared the ratio of current smokers across the patient groups in years 2010–2011 (N = 2043/5885, 34.7%) and 2015–2016 (N = 1962/6411, 30.6%), and found a 4.1% decline in active smoking.

Table 1. Characteristics and smoking statuses of the patient groups studied based on the two-year follow-up.

	All N 35650	Asthma N 4 549	COPD[a] N 2 111	Sleep apnoea N 5 931	Type 1 diabetes N 632	Type 2 diabetes N 8 281	IHD[2] N 9 200	Cerebral infarction N 4 946
Men	18,997 (53.3)	1 455 (32.0)	1 397 (66.2)	3 945 (66.5)	335 (53.0)	4 267 (51.5)	5 246 (57.0)	2 352 (47.6)
Women	16,653 (46.7)	3 094 (68.0)	714 (33.8)	1 986 (33.5)	297 (47.0)	4 014 (48.5)	3 954 (43.0)	2 594(52.4)
Mean age (SD)	63.5 (16.4)	50.7 (19.0)	66.3 (10.8)	53.5 (12.7)	40.6 (18.3)	65.9 (13.4)	71.2 (13.0)	70.8 (14.2)
N of patients with at least one documented smoking status in EHR [3]	21,372 (59.9)	2775(61.0)	1820(86.1)	4 949(83.4)	331 (52.4)	3 634 (43.9)	5 005(54.4)	2858(57.8)
Smoking status[4]								
Current smoker	7 105/21,372 (33.2)	813/2775 (29.3)	1 268/1820 (69.7)	1 243/4 949 (25.1)	118/331 (35.6)	1135/3 634 (31.2)	1 586/5 005 (31.7)	942/2 585 (33.0)
Ex-smoker	4 852/21,372 (22.7)	599/2775 (21.6)	501/1820 (27.5)	1 384/4 949 (28.0)	38/331(11.5)	820/3 634 (22.6)	1 114/5 005 (22.3)	396/2 585 (13.9)
Never smoker	9 415/21,372 (44.1)	1363/2775 (49.1)	51/1820(2.8)	2 322/4 949 (46.9)	175/331 (52.9)	1 679/3 634 (46.2)	2 305/5 005 (46.1)	1520/2 585 (53.2)
N of texts per patient including smoking status, mean (SD)[5]	2.0(3.2)	1.7(2.7)	5.3 (5.8)	1.8(2.3)	1.5(3.3)	1.3(2.7)	1.8 (2.9)	2.4(3.6)
N of events per patient during the follow-up, median (IQR)[5]	18(9–34)	14(6–28)	22(10–43)	12[7–22]	18(8–31)	17(8–35)	19(11–34)	28(17–45)

Data is presented as n (%) unless otherwise stated. SD = Standard Deviation. N = Number. IQR = Inter quartile range.
[a]Chronic obstructive pulmonary diseases [2]Ischemic heart diseases [3]Electronic health records [4] percentages have been calculated from the patients whose EHR contains at least one smoking related sentence. [5]One event/text = one visit

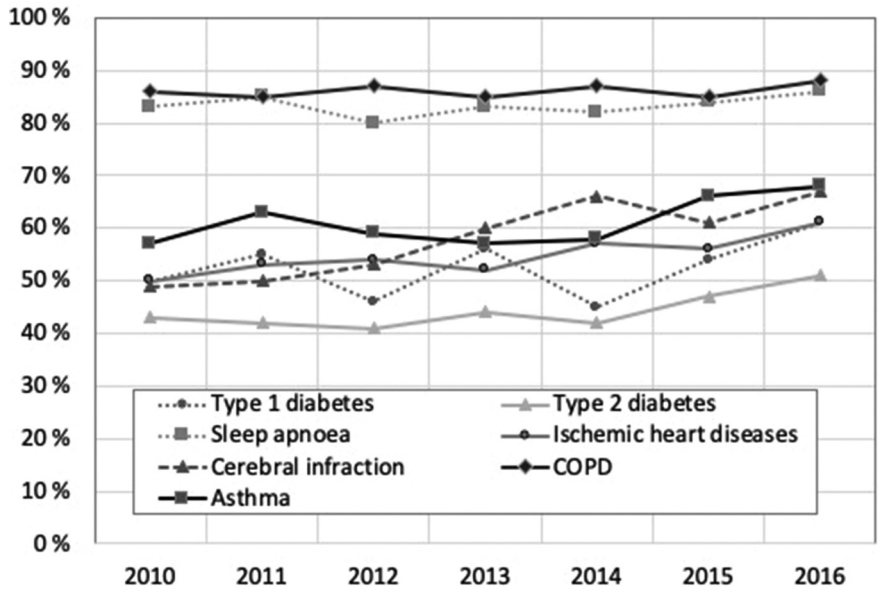

Figure1. Percentage of patients with at least one documented smoking status during a two-year follow-up.

Smoking cessation

About half of the current smokers (49%) had discussed smoking cessation with the healthcare professional (Table 2). When the currently smoking asthma patients were selected as the control group, smoking cessation was discussed 9% more frequently with COPD patients and 11% more frequently with type 1 diabetes patients. On few occasions had current smokers been referred to the nurse-managed smoking cessation programme. However, asthma, COPD and sleep apnoea patients had significantly more visits to the intervention nurse than the other patient groups (223/3324, 6.7% vs 55/3781, 1.5%, $p < 0.001$).

Validation of the algorithms used

The performance of the rule-based algorithm 1 in finding smoking-related sentences is shown in Table 3. The ULMFiT-based classifier performed best for never smokers (F1-score 91.9). For the classification of ex-smokers and current smokers, the F1-scores were 80.4 and 78.5, respectively. The errors observed were related [1] to the differentiation between current and former smoker [2], to ambiguous expressions such as 'the patient has a long smoking history' and 'the patient is an occasional smoker', and [3] to exposure to passive smoking. Among current smokers, an F1-score of the rule-based algorithm 2 (identifying sentences related to smoking cessation interventions) was 87.9.

Table 2. Encouragement and active intervention for smoking cessation during a two-year follow-up among the current smokers.

	All N 7 105	Asthma N 813	COPD[1] N 1 268	Sleep apnoea N 1 243	Type 1 diabetes N 118	Type 2 diabetes N 1 135	IHD[2] N 1 586	Cerebral infarction N 942
Smoking cessation encouraged by the clinician	3470 (48.8)	450 (55.4)	755 (59.5)	609 (49.0)	72 (61.0)	498 (43.9)	650 (41.0)	436 (46.3)
differences among subgroups relative to the asthma group	-	1	1.09	0.89	1.11	0.8	0.75	0.84
referred to nurse – managed smoking cessation program	278 (3.9)	74 (9.1)	89 (7.0)	60 (4.8)	1 (0.8)	23 (2.0)	23 (1.5)	8 (0.8)

Data is presented as n (%) unless otherwise stated. [1] Chronic obstructive pulmonary diseases [2]Ischemic heart diseases

Table 3. Performance of the algorithms employed.

Tested algorithm	Performance of the algorithm	Accuracy	Precision	Recall	F1-score
rule-based 1	in identifying smoking related sentences	94.3	99.0	93.3	96.1
ULMFiT-based language model	in classifying smoking statuses current smoker	85.9	66.2	96.2	78.5
	ex-smoker	89.9	97.6	68.3	80.4
	never smoker	93.4	94.9	89.2	91.9
rule-based 2	in identifying sentences related to smoking cessation interventions	87.0	94.0	82.5	87.9

Discussion

In the present registry-based study, we report how often a patient's smoking status and the conversation between the patient and the clinician about smoking cessation were documented in the narrative reports of EHR. A large number of reports of the patients with asthma, COPD, sleep apnoea, type 1 and 2 diabetes, IHD and cerebral infarction were followed for two years starting from their first visit to the Turku University Hospital due to the foregoing disease. We used a combination of rule-based and deep learning-based algorithms to extract and classify smoking statuses from written language of EHR. In 60% of patients, we found smoking status documented. On average, smoking status was documented significantly more frequently among the COPD (86%) and the sleep apnoea (83%) patients than in the rest of the patient groups (53%). Half of the patients classified as current smokers had discussed smoking cessation with the clinician. The trends over 9-year observation period showed that clinicians documented smoking status more often in years 2016–2018 than in years 2010–2012.

The Finnish EHRs include both structured and non-structured elements where smoking status is often documented in a non-structured manner. Many clinicians believe point-and-click EHR templates can limit their ability to capture the unique clinical story and to adequately document their medical decision-making process which is unique to each patient encounter [17]. On the other hand, without validated language models and classifiers working in multiple languages, narrative reports might become a challenge for building the foundation for evidence-based medicine and clinical decision support needed in every hospital. In this study, we showed that the performance of the ULMFiT-based algorithm varied from good to excellent in classifying patients' smoking status from Finnish narrative reports. Overall, ULMFiT and other deep learning-based approaches have shown to be promising tools in standardization of language used in narrative reports including abbreviations, acronyms, eponyms, slang and jargon words [18,19]. An additional benefit of using this type of language model is that, once fine-tuned for Finnish medical narratives, the classifier can be further developed for other study needs.

Finnish clinical guidelines encourage the clinicians to ask patients about smoking and advise the smokers to quit [20]. In Finland, 14% of people smoke daily [2]. The overall decline in active smoking during the study years (4%) was similar what has been reported in general adult population [2]. It is possible that documentation of smoking is most often missed when the patient does not smoke, since then no intervention is needed. Prior studies have reported that 44–95% of patients with asthma, COPD or diabetes have a smoking status documented in primary care EHR [21–24]. Studies in secondary health care are scarce. Our study showed that clinicians ask smoking more frequently during recent years compared to before. The trends improved especially in the patient groups with the poorest documentation in the beginning of the study period. Due to the retrospective study design, the physicians or patients were not aware of the study, which increases the reliability of the results. Smoking status was documented most actively in patients with sleep apnoea and COPD. In the study hospital, asthma, sleep apnoea and COPD patients were treated by pulmonologists. It did not, however, explain the observed differences alone. Compared to other specialities, a pulmonologist may ask about smoking more systematically and the implementation of preliminary information forms some years ago have probably increased documentation activity. A recent study also found, that pulmonologists experience less barriers, such as lack of training, than the other specialists [25]. In the present study, the minor proportion of patients had participated in the nurse-managed smoking cessation program. That is most likely due to that fact that the Finnish primary healthcare system has the main responsibility in counselling and managing smoking cessation programs. However, physicians working in secondary care should also use their authoritative role in supporting cessation [5]. As a limitation, our study included patients only from one hospital and patients were classified only to one disease group based on the diagnosis that appeared first. This choice was made on the basis of making the patient group definitions and follow-up time definitions unique and simple. A more refined approach could be used in future studies

Previous studies have shown that, although physicians ask about smoking, they are less likely to offer practical advice to quit [25]. In our study, 49% of current smokers had discussed smoking cessation with the physician. In addition to behavioural support, pharmacological treatment has also shown to increase the success rates in smoking cessation [26]. First-line pharmacotherapies for smoking cessation include nicotine replacement therapy (NRT). Since NRT is based on over-the-counter products, it was unfortunately impossible for us to follow the treatment through hospital EHR. It is also possible that clinicians discussed smoking and smoking cessation with the patients more active than what was documented in EHR. The highest

intervention rate was among patients with type 1 diabetes (61%). These patients were younger than those in the other disease groups that may affect the physician's activity. However, it seems that the clinicians still miss opportunities to talk about the importance of smoking cessation on long-term outcomes. Prior studies have found that the knowledge, attitudes, interest, lack of time and confidence are the common reasons not to implement smoking cessation intervention more effectively [5,25]. However, many specialists do not refer smokers to a cessation nurse either [25]. Interestingly, in our study, the proportion sleep apnoea patients who were currently smoking and were encouraged to quit smoking by the clinician, was 11% less compared to asthma patients despite more frequent documentation of smoking status. The reason for this finding is unclear. One could speculate that clinicians do not consider smoking in sleep apnoea patients as harmful as in asthma patients. Another reason might be that sleep apnoea patients are less likely to accept referral to a smoking cessation program due to the fear of gaining more weight after smoking cessation. Furthermore, patients' interest, physicians' unawareness of available services and disregard for shared responsibilities could explain the discrepancy.

Conclusion

In conclusion, even when the negative effects of smoking on treatment outcomes are well established, physicians still do not systematically document patients' smoking status. Therefore, it is possible that they do not take potential smoking into account when monitoring therapy outcomes either. Secondly, EHRs are growing sources for real-world data studies increasing the need for natural language processing. In the present study, we showed that a deep learning-based ULMFiT classifier can detect and classify patients' smoking status efficiently from medical narrative reports.

Abbreviations

EHR=electronic health records, NLP= natural language processing, ULMFiT=Universal Language Model Fine-tuning, COPD=chronic obstructive pulmonary disease, IHD=ischemic heart diseases.

Acknowledgments

The authors would like to thank Mikko Tukiainen for his involvement in the development of the ULMFiT algorithm and Emily Kemp for proofreading the article.

Disclosure statement

No potential conflict of interest was reported by the author (s).

Funding

This work was supported by The Research Foundation of the Pulmonary Diseases; Väinö ja Laina Kivi Foundation under grant Väinö ja Laina Kiven Säätiö 202000069 and Turku University Hospital VTR funding.

ORCID

Eveliina Hirvonen http://orcid.org/0000-0002-4921-3387

References

[1] World Health Organization. WHO REPORT On The Global Tobacco Epidemic, 2015 raising taxes on tobacco. 2015 [cited 2017 Jan 22]; Available from: https://www.who.int/health-topics/tobacco

[2] Virtanen S, Jääskeläinen M. Tobacco statistics 2018. 2019; Available from: http://urn.fi/URN:NBN:fi-fe2019121046603

[3] US Department of Health and Human Services. Smoking Cessation: A Report of the Surgeon General. Atlanta, GA; 2020.

[4] Maddatu J, Anderson-Baucum E, Evans-Molina C. Smoking and the risk of type 2 diabetes [Internet]. Vol. 184, Translational Research. Mosby Inc.; 2017 [cited 2020 Oct 22]. p. 101–7. Available from: https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/28336465/

[5] Keto J, Jokelainen J, Timonen M, et al. Physicians discuss the risks of smoking with their patients, but seldom offer practical cessation support. Subst Abuse Treat Prev Policy [Internet]. 2015 Dec 2 [cited 2019 Sep 17];10 (1):43. Available from:: http://substanceabusepolicy.biomedcentral.com/articles/10.1186/s13011-015-0039-9

[6] Jamal A, Dube SR, King BA. Tobacco use screening and counseling during hospital outpatient visits among US adults, 2005-2010. Prev Chronic Dis [Internet]. 2015 cited 2020 Nov 4;12(8). Available from. https://pubmed.ncbi.nlm.nih.gov/26292063/

[7] West R, Raw M, McNeill A, et al. Health-care interventions to promote and assist tobacco cessation: a review of efficacy, effectiveness and affordability for use in national guideline development. Addiction [Internet]. 2015 Sep 1 [cited 2020 Nov 4];110(9):1388–1403. Available from: https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/26031929/

[8] Stead LF, Buitrago D, Preciado N, et al. Physician advice for smoking cessation [Internet]. Vol. 2017, Cochrane Database of Systematic Reviews. John Wiley and Sons Ltd; 2013 [cited 2021 May 22]. Available from: https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/23728631/

[9] Hartmann-Boyce J, Hong B, Livingstone-Banks J, et al. Additional behavioural support as an adjunct to pharmacotherapy for smoking cessation [Internet]. Vol. 2019, Cochrane Database of Systematic Reviews. John

Wiley and Sons Ltd; 2019 [cited 2021 May 22]. Available from: https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/31166007/

[10] Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol [Internet]. 2020 Feb 1 [cited 2020 Oct 22];145(2):463–469. Available from: https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/31883846/

[11] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–1359.

[12] Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018 [cited 2020 Apr 23]; Available from: http://nlp.fast.ai/ulmfit.

[13] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv [Internet]. 2018 Oct 10 [cited 2020 Nov 17];4171–4186. Available from: http://arxiv.org/abs/1810.04805

[14] Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv [Internet]. 2019 Jul 26 [cited 2020 Nov 17]; Available from: http://arxiv.org/abs/1907.11692

[15] Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. arxiv.org [Internet]. 2020 May 28 [cited 2020 Nov 17]; Available from: http://arxiv.org/abs/2005.14165

[16] Karlsson A, Ellonen A, and Irjala H, et al. Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit. In: ESMO OPEN 6(3); 2021 Jun 100175 .

[17] Barry J. Value of unstructured patient narratives. Health Manag Technol [Internet]. 2010 [cited 2020 Dec 16];31(7):6–7. Available from: https://pubmed.ncbi.nlm.nih.gov/20712268/

[18] Syed K, Sleeman W, Hagan M, et al. Automatic incident triage in radiation oncology incident learning system. Healthcare [Internet]. 2020 Aug 14 [cited 2020 Nov 6];8(3):272. Available from:: /pmc/articles/PMC7551126/?report=abstract

[19] Swaminathan KK, Mendonca E, Mukherjee P, et al. Development of an algorithm using natural language processing to identify metastatic breast cancer patients from clinical notes. J Clin Oncol. 2020 May 20;38(15_suppl):e14056–e14056.

[20] Working group set up by the Finnish Medical Society Duodecim and The Finnish Association for General Practice. Tobacco and nicotine dependency, prevention and treatment. Current Care Guidelines [Internet]. Helsinki; 2018. Available from: www.kaypahoito.fi

[21] Heinmüller S, Schaubroeck E, Frank L, et al. The quality of COPD care in German general practice—A cross-sectional study. Chron Respir Dis [Internet]. 2020 cited 2021 May 14; 17. Available from:. https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/33272029/

[22] Lange P, Rasmussen FV, Borgeskov H, et al. The quality of COPD care in general practice in Denmark: The KVASIMODO study. Prim Care Respir J [Internet]. 2007Jun cited 2021 May 14;16(3):174–181. Available from. https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/17516009/

[23] Bailey SR, Fankhauser K, Marino M, et al. Smoking assessment and current smoking status among adolescents in primary care settings. Nicotine Tob Res [Internet]. 2020 Nov1 [cited 2021 May 14];22(11):2098–2103. Available from:: https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/32556337/

[24] Kaufmann C, Markun S, Hasler S, et al. Performance measures in the management of chronic obstructive pulmonary disease in primary care - A retrospective analysis. Praxis (Bern 1994) [Internet]. 2015 Jan 1 [cited 2021 May 14];104(17):897–907. Available from:: https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/26286494/

[25] Meijer E, Van Der Kleij RMJJ, Chavannes NH. Facilitating smoking cessation in patients who smoke: a large-scale cross-sectional comparison of fourteen groups of healthcare providers. BMC Health Serv Res [Internet]. 2019Oct25 cited 2020 Nov 6;19(1). Available from:. https://pubmed-ncbi-nlm-nih-gov.ezproxy.utu.fi/31653215/

[26] Fant RV, Buchhalter AR, Buchman AC, et al. Pharmacotherapy for tobacco dependence. Handb Exp Pharmacol [Internet]. 2009 [cited 2020 Dec 22];192(192):487–510. Available from:: https://pubmed.ncbi.nlm.nih.gov/19184660/