


# Smoking is a predictor of complications in all types of surgery: a machine learning-based big data study

Helene L. Gräsbeck<sup>1,2,\*</sup> , Aleksi R. P. Reito<sup>3,4</sup>, Heikki J. Ekroos<sup>5</sup>, Juhani A. Aakko<sup>6</sup>, Olivia Hölsä<sup>6</sup> and Tuula M. Vasankari<sup>7,8</sup>

<sup>1</sup>Department of Pulmonary Medicine, HUS Heart and Lung Center, Helsinki, Finland

<sup>2</sup>Doctoral Programme of Clinical Research, University of Helsinki, Helsinki, Finland

<sup>3</sup>Center for Musculoskeletal Diseases, Tampere University Hospital, Tampere, Finland

<sup>4</sup>Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

<sup>5</sup>Pulmonary Unit, HUS Porvoo Hospital, Porvoo, Finland

<sup>6</sup>Medaffcon Oy, Espoo, Finland

<sup>7</sup>Department of Pulmonary Diseases and Clinical Allergology, University of Turku, Turku, Finland

<sup>8</sup>Finnish Lung Health Association (Filha), Helsinki, Finland

\*Correspondence to: Helene L. Gräsbeck, Department of Pulmonary Medicine, HUS Heart and Lung Center, PO Box 340, 00029 HUS, Helsinki, Finland (e-mail: helene.grasbeck@hus.fi)

This paper is based on a communication to a meeting at the Department of Pulmonary Medicine, Meilahti Triangle Hospital, Helsinki University Central Hospital on 28 March 2022, and a poster presentation at the Nordic Lung Congress in Copenhagen on 1–3 June 2022.

## Abstract

**Background:** Machine learning algorithms are promising tools for smoking status classification in big patient data sets. Smoking is a risk factor for postoperative complications in major surgery. Whether this applies to all surgery is unknown. The aims of this retrospective cohort study were to develop a machine learning algorithm for clinical record-based smoking status classification and to determine whether smoking and former smoking predict complications in all surgery types.

**Methods:** All surgeries performed in a Finnish hospital district from 1 January 2015 to 31 December 2019 were analysed. Exclusion criteria were age below 16 years, unknown smoking status, and unknown ASA class. A machine learning algorithm was developed for smoking status classification. The primary outcome was 90-day overall postoperative complications in all surgeries. Secondary outcomes were 90-day overall complications in specialties with over 10 000 surgeries and critical complications in all surgeries.

**Results:** The machine learning algorithm had precisions of 0.958 for current smokers, 0.974 for ex-smokers, and 0.95 for never-smokers. The sample included 158 638 surgeries. In adjusted logistic regression analyses, smokers had increased odds of overall complications (odds ratio 1.17; 95 per cent c.i. 1.14 to 1.20) and critical complications (odds ratio 1.21; 95 per cent c.i. 1.14 to 1.29). Corresponding odds ratios of ex-smokers were 1.09 (95 per cent c.i. 1.06 to 1.13) and 1.09 (95 per cent c.i. 1.02 to 1.17). Smokers had increased odds of overall complications in all specialties with over 10 000 surgeries. ASA class was the most important complication predictor.

**Conclusion:** Machine learning algorithms are feasible for smoking status classification in big surgical data sets. Current and former smoking predict complications in all surgery types.

## Introduction

Globally, tobacco use causes more than 8 million deaths per year and is the leading preventable cause of death<sup>1</sup>. A recent American, state-wide study found that 24.1 per cent of surgical patients smoked cigarettes. This is higher than the national average in the general population<sup>2</sup>.

Smoking increases the risk of surgical complications through several mechanisms. It promotes atherosclerosis by altering the lipid profile, damaging the vascular endothelium, and increasing oxidative stress, neutrophil count, and hypercoagulability<sup>3</sup>. Further, cigarette smoke components impair wound healing, thereby increasing the risk of wound dehiscence and infection<sup>4,5</sup>.

In 2014, a comprehensive meta-analysis of 100 cohort and seven case-control studies found an elevated risk of general morbidity rate, intensive care admission, general infections, and

pulmonary, wound, and neurological complications in smokers undergoing various types of surgery<sup>6</sup>. A narrative review conducted in 2015 found consistent evidence, stressing the risk of postoperative wound healing-related and cardiovascular complications in smokers<sup>7</sup>.

However, most studies on smoking-related postoperative complications are small and, of the cohort studies included in the 2014 meta-analysis, only two had a study population of over 100 000. Most of the studies only included one or few surgical specialties, and the large-volume specialty gynaecology was not covered in any of the studies<sup>6</sup>.

When studying large data sets, determining smoking status based on electronic health record (EHR) free-text notes often forms a major obstacle. The manual extraction of smoking status from EHRs is both laborious and expensive. Natural

Received: November 29, 2022. Accepted: January 22, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of BJS Society Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

language processing (NLP) methods have proven feasible for clinical text classification tasks<sup>8</sup>. Furthermore, some recent studies have demonstrated that machine learning-based NLP methods are accurate in smoking status extraction from EHR notes<sup>9–11</sup>.

The aims of this study were to develop a well-functioning MLA for the classification of smoking status in a large patient population and to use this to assess smoking and former smoking as risk factors for postoperative complications in all types of inpatient and outpatient surgery, utilizing a large real-life sample. In addition, a further aim was to clarify the global importance of smoking as a risk factor for surgical complications in relation to other known risk factors.

## Methods

### Study sample and data source

This was a retrospective cohort study. The data comprised all surgical procedures performed between 1 January 2015 and 31 December 2019 in the Finnish hospital district of Helsingin ja Uudenmaan Sairaanhoidopiiri (HUS). HUS is a specialized healthcare provider and the largest hospital district in Finland, with a population base of roughly 1.7 million. The data source was HUS Datalake, a public database service for clinical researchers. HUS Datalake contains pseudonymized information from the local EHR systems, such as diagnosis codes, free-text clinical notes, and clinical measurement results. The patients were identified with the aid of Opera, an electronic notification created for each patient undergoing a procedure. Exclusion criteria were age below 16 years, non-surgical procedure, unknown smoking status, and unknown ASA class.

### Outcomes and explanatory variables

The primary outcome was the composite of any perioperative or postoperative 90-day complication in all surgeries. ICD-10 diagnosis codes for wound, cardiovascular, neurological, respiratory, thromboembolic, gastroenterological, urinary, and orthopaedic complications and unspecified bacterial infections, and the occurrence of death, reoperation, hospital readmission, and mechanical ventilation or ICU admission were included in the outcome composite (Table S1). Any of these events recorded within 90 days after surgery was considered a complication. Secondary outcomes were any 90-day complication in specialties with more than 10 000 operations (gastroenterological, orthopaedic, gynaecological, plastic, and otorhinolaryngological surgery), and, in all surgeries, critical 90-day grade IV–V (life-threatening complications and death) complications as defined in the Clavien–Dindo classification<sup>12</sup>. These critical complications were acute coronary syndrome, shock, cerebral infarction, pulmonary embolism, peritonitis, mechanical ventilation or ICU admission, and death. Severe complications of grade III (requiring re-intervention) were not included in this secondary analysis, as this was a registry study and determination of individual patient paths would not have been possible.

Demographic data included age and sex. Patient characteristics included preoperative smoking status, ASA class, and chronic diseases as classified in the Charlson co-morbidity index (CCI). CCI was calculated using ICD-10 codes, as implemented in the co-morbidity R package<sup>13,14</sup>.

### Definition of smoking status and training of the machine learning algorithm

EHR free-text sentences containing the word stems tupak\* ('tupakka' meaning tobacco in Finnish and 'tupakointi' meaning smoking), aski\* ('aski' meaning pack), and smok\* were utilized for smoking status determination. For this purpose, all EHR free-text notes recorded by physicians between 2014 and 2019, including all medical specialties, were retrieved from HUS Datalake. A random sample of sentences were annotated manually into smoking status classes for training and testing of the MLA. Two medical doctors (H.L.G. and contributor J.N.) independently annotated the sentences according to predefined rules. According to West<sup>15</sup>, long-term smoking abstinence has often referred to abstinence for at least 6 months. Applying this definition, a current smoker had smoked any amount of cigarettes within the previous 6 months, an ex-smoker had stopped smoking at least 6 months previously, and a never-smoker had never smoked. If the smoking status could not be determined, it was annotated as unknown. Users of electronic cigarettes were assigned as unknown smoking status. The sentences were annotated in the HUS Datalake analysis environment.

The annotated sentences were converted into lower case and all special characters were removed. The sentences were split into training data (90 per cent of the annotated sentences) and test data (10 per cent of the annotated sentences). A supervised machine learning-based text classifier fastText<sup>16</sup> was applied to classify smoking-related sentences by smoking status into the classes ex-smoker, never-smoker, current smoker, and unknown. An unknown class was used to filter unclear and uncertain sentences. Moreover, the classified sentences with low classification probabilities were assigned to the unknown class, where the thresholds for the classification probabilities were tuned in favour of specificity. MLA training and assessment of the MLA was performed using five-fold cross-validation.

Finally, surgery-level classification was performed by assigning a smoking status of ex-smoker, never-smoker, or current smoker to the patient if smoking-related sentences classified into one of those three classes were found in the medical record. If multiple smoking statuses were available for a patient prior to surgery and they were conflicting, the status of the most recent classification was assigned to the surgery, unless the status was a never-smoker, in which case the smoking status assigned to the surgery was an ex-smoker. Further details of the MLA training process and performance have been published elsewhere<sup>17</sup>.

### Statistical analysis

The impact of smoking status (never-smoker, ex-smoker, and current smoker) on postoperative complications was assessed using binary logistic regression unadjusted and adjusted with age as a continuous covariate and sex, CCI, and ASA class as categorical covariates. Surgeries with missing covariate values were omitted. Analyses were performed for all surgeries concerning overall complications and critical complications, and independently for specialties with more than 10 000 operations concerning overall complications. Corresponding unadjusted and adjusted ORs with 95 per cent c.i. were calculated. The relative importance of the variables in the model was assessed by calculating the partial chi-squared statistic for each variable, as well as a global importance score based on Shapley values<sup>18</sup>, which were calculated as implemented in the fastshap R package<sup>19</sup>. Nagelkerke's pseudo R<sup>2</sup> and McFadden's pseudo R<sup>2</sup> were computed to assess the overall model fit.

Data processing, statistical analysis, and modelling were performed using Python (version 3.8.4)<sup>20</sup> and R (version 3.6.3)<sup>21</sup>. All relevant data are within the paper and its [Supplementary material](#).

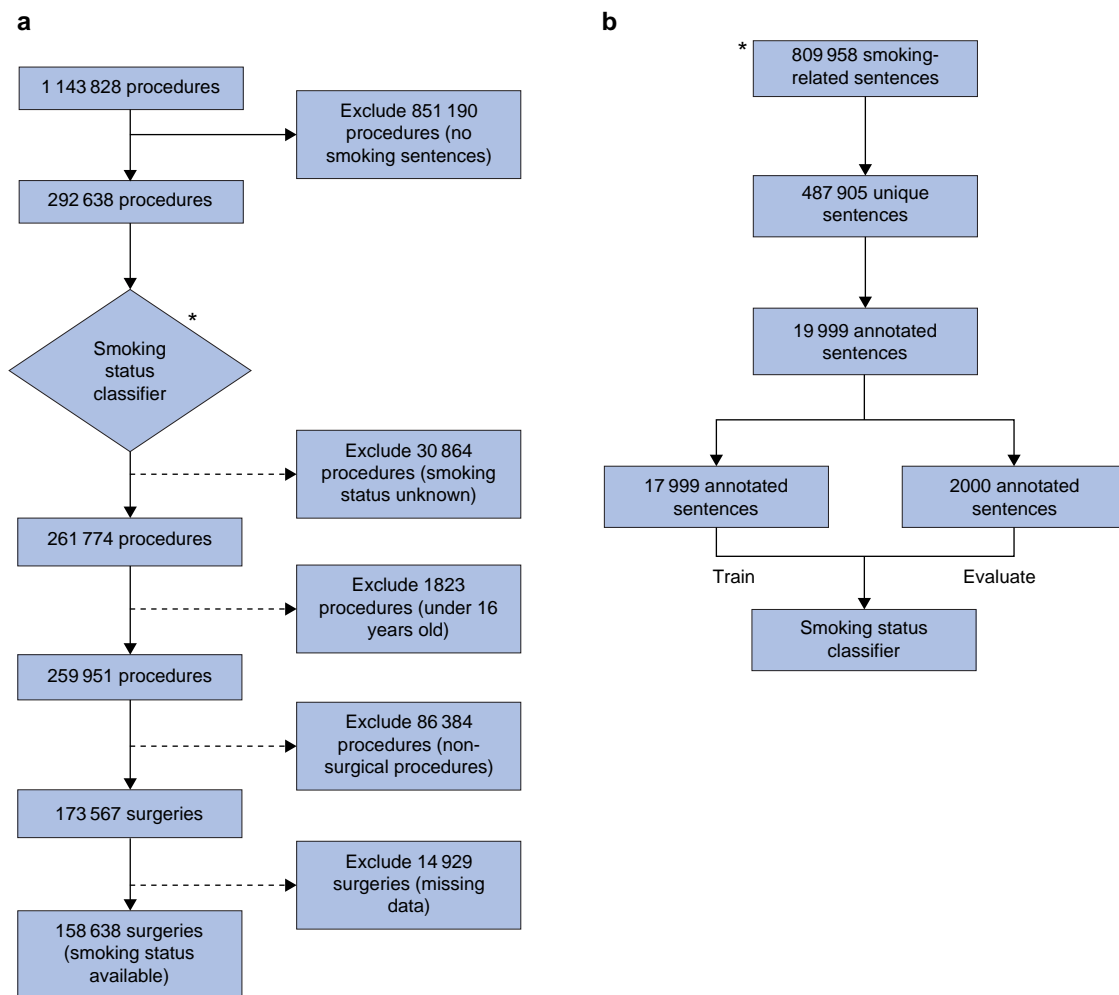
## Results

Roughly 500 000 unique sentences were retrieved from HUS Datalake, out of which 19 999 were randomly sampled and annotated manually into smoking status classes for training and testing of the MLA (Fig. 1). Initially, all sentences were classified into classes: ex-smoker, current smoker, never-smoker, and unknown. After tuning the threshold for uncertain classifications, the overall accuracy was 79.0 per cent (95 per cent c.i., 77.3 to 80.8 per cent) with precision of 95.8 per cent (95 per cent c.i., 93.5 to 98.1 per cent), 97.4 per cent (95 per cent c.i. 95.8 to 98.9 per cent), 95.0 per cent (95 per cent c.i. 93.5 to 96.6 per cent), and 55.8 per cent (95 per cent c.i. 51.5 to 60.1 per cent), and recalls of 68.7 per cent (95 per cent c.i. 63.4 to 73.9 per cent), 93.2 per cent (95 per cent c.i. 90.7 to 95.7 per cent), 67.6 per cent (95 per cent c.i. 64.4 to 70.9 per cent), and 92.0 per cent (95 per cent c.i. 89.6 to 94.3 per cent) for the ex-smoker, never-smoker, current smoker, and unknown classes respectively. When excluding the unknown class

and assessing the performance of the MLA separately for the classes ex-smoker, never-smoker, and current smoker, the average precision and recall were 96.1 per cent and 84.6 per cent respectively.

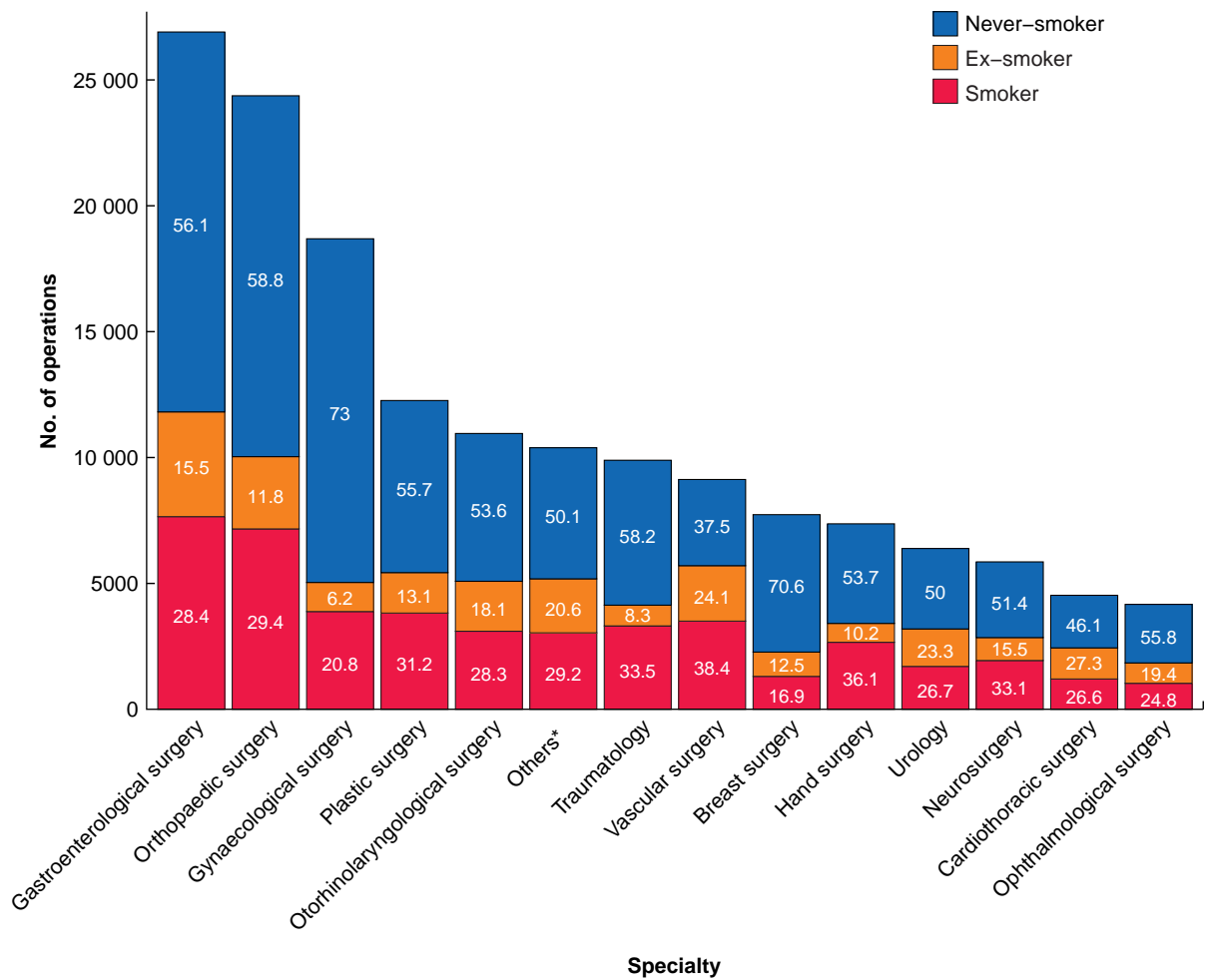
Smoking-related sentences were found in 292 638 of 1 143 828 procedures (Fig. 1). Smoking status was assigned to 261 774 procedures using the trained MLA. The final sample included 158 638 surgeries within 19 surgical specialties (Fig. 2), after removing non-surgical procedures, procedures with missing data, and patients under 16 years of age. The proportions of current smokers in the surgical specialties varied between 16.9 per cent in breast surgery (1308 of 7732 patients) and 38.4 per cent in vascular surgery (3503 of 9128 patients) (Fig. 2). The ratio of current smokers to the number of surgeries showed a downward trend during the study interval, from 33 per cent (6639 of the total of 20 435 surgeries) in 2015 compared with 26 per cent (9518 of the total of 36 488 surgeries) in 2019 (Fig. 3).

The patients' baseline characteristics and demographics are displayed in Table 1. Ex-smokers were the oldest, with a median age of 66 (range 56–74) years, compared with 55 (range 41–65) years in current smokers and 57 (range 41–70) years in never-smokers. Ex-smokers were the most likely to be male and had the largest co-morbidity burden, with a large proportion in



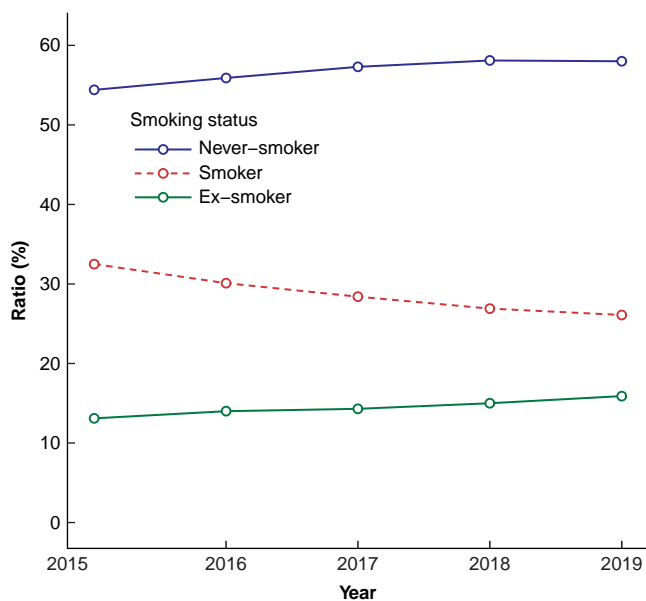
**Fig. 1** Sample selection and development of the machine learning algorithm

**a** Study flow diagram displaying the total number of procedures in the retrieved data, the number of procedures excluded at each step of the selection process, and the number of surgeries included in the final sample. **b** A proportion of all extracted unique smoking-related sentences were annotated manually into smoking status classes for training and testing of the machine learning algorithm.



**Fig. 2 Surgical specialties: numbers of operations and smoking status class proportions**

Surgical specialties included in the final sample, with their numbers of operations, and numbers of never-smokers, ex-smokers, and current smokers. Percentages are shown within the boxes. \*Includes thoracic surgery, endocrine surgery, general surgery, liver surgery, and transplant surgery.



**Fig. 3 Surgical patients' smoking status ratios in the study years 2015–2019**

ASA class III or IV and having numerous diseases included in the CCI. They were also the most likely to have any of the potentially smoking-related CCI diseases.

The overall complication rate was 56 690 (35.7 per cent) in all surgeries. Corresponding figures were 9921 of 26 903 (36.9 per cent) in gastroenterological surgery, 6364 of 24 372 (26.1 per cent) in orthopaedic surgery, 4982 of 18 691 (26.7 per cent) in gynaecological surgery, 5523 of 12 265 (45.0 per cent) in plastic surgery, and 2440 of 10 955 (22.3 per cent) in otorhinolaryngological surgery. The critical complication rate was 6925 (4.4 per cent) in all surgeries. Logistic regression analysis of all surgeries showed that both ex-smokers and current smokers had increased odds of overall postoperative complications compared with never-smokers (Table 2). The same was true after multivariable adjustment, with OR 1.09 (95 per cent c.i. 1.06 to 1.13) for ex-smokers and OR 1.17 (95 per cent c.i. 1.14 to 1.20) for current smokers. Both ex-smokers and current smokers had increased unadjusted overall complication odds in all secondary analyses. After adjustment for covariates, odds were significantly increased for ex-smokers in gastroenterological surgery (OR 1.09; 95 per cent c.i. 1.01 to 1.18), plastic surgery (OR 1.24; 95 per cent c.i. 1.10 to 1.40), and otorhinolaryngological surgery (OR 1.20; 95 per cent c.i. 1.05 to 1.37). For current smokers, odds were significantly increased in

Table 1 Baseline characteristics

Characteristic	Never-smoker, n = 90 208	Ex-smoker, n = 23 101	Current smoker, n = 45 326
<b>Sex</b>			
Male	32 199	13 586	23 290
Female	58 009	9515	22 036
<b>Age (years)</b>			
Mean(s.d.)	56(18)	64(14)	53(16)
Median (i.q.r.)	57 (41–70)	66 (56–74)	55 (41–65)
<b>ASA class*</b>			
1	19 656 (21.8)	1551 (6.7)	4977 (11.0)
2	37 924 (42.0)	6911 (29.9)	18 365 (40.5)
3	25 126 (27.9)	9 914 (42.9)	15 573 (34.4)
4	6964 (7.7)	4449 (19.3)	5828 (12.9)
5	538 (0.6)	276 (1.2)	583 (1.3)
<b>CCI*</b>			
0†	54 177 (60.1)	7860 (34.0)	24 734 (54.6)
1†	10 750 (11.9)	4328 (18.7)	7621 (16.8)
2†	16 595 (18.4)	5311 (23.0)	7007 (15.5)
3†	4064 (4.5)	2336 (10.1)	2851 (6.3)
4†	1638 (1.8)	1128 (4.9)	1265 (2.8)
≥5†	2984 (3.3)	2138 (9.3)	1848 (4.1)
<b>Acute myocardial infarction‡</b>	2058 (2.3)	1465 (6.3)	1602 (3.5)
<b>Congestive heart failure‡</b>	2468 (2.7)	1784 (7.7)	1374 (3.0)
<b>Peripheral vascular disease‡</b>	3944 (4.4)	3470 (15.0)	4859 (10.7)
<b>Cerebrovascular disease‡</b>	5640 (6.3)	2692 (11.7)	3571 (7.9)
<b>Chronic pulmonary disease‡</b>	4829 (5.4)	3742 (16.2)	4393 (9.7)
<b>Hemiplegia or paraplegia‡</b>	272 (0.3)	94 (0.4)	198 (0.4)
<b>Cancer (any malignancy)‡</b>	18 465 (20.5)	6892 (29.8)	7636 (16.8)
<b>Metastatic solid tumour‡</b>	1260 (1.4)	566 (2.5)	661 (1.5)
<b>Other co-morbidity</b>	5035 (5.6)	2401 (10.4)	3566 (7.9)

Values are n (%) unless otherwise indicated. \*Percentages have been rounded and might not total 100. †Total number of CCI co-morbidities. ‡Potentially smoking-related CCI co-morbidities. CCI, Charlson co-morbidity index.

Table 2 Complication rates and association between smoking status and outcomes

Outcome	Specialty	Smoking status		
		Never-smoker	Ex-smoker	Current smoker
Any 90-day complication	All surgeries	27 994 (31.0)	10 401 (45.0)	18 295 (40.4)
	Unadjusted	1 (Reference)	1.82 (1.77 to 1.87)*	1.5 (1.47 to 1.54)*
	Adjusted	1 (Reference)	1.09 (1.06 to 1.13)*	1.17 (1.14 to 1.20)*
	Gastro	4902 (32.5)	1847 (44.4)	3172 (41.4)
	Unadjusted	1 (Reference)	1.66 (1.55 to 1.78)*	1.47 (1.39 to 1.56)*
	Adjusted	1 (Reference)	1.09 (1.01 to 1.18)*	1.21 (1.13 to 1.29)*
	Ortho	3328 (23.2)	893 (31.1)	2143 (29.9)
	Unadjusted	1 (Reference)	1.49 (1.37 to 1.63)*	1.41 (1.32 to 1.5)*
	Adjusted	1 (Reference)	1.01 (0.92 to 1.11)*	1.36 (1.27 to 1.46)*
	GYN	3477 (25.5)	357 (30.9)	1148 (29.6)
	Unadjusted	1 (Reference)	1.31 (1.15 to 1.49)*	1.23 (1.14 to 1.33)*
	Adjusted	1 (Reference)	1.12 (0.98 to 1.28)*	1.12 (1.03–1.21)*
	Plastics	2647 (38.7)	868 (54.0)	2008 (52.5)
	Unadjusted	1 (Reference)	1.86 (1.67 to 2.08)*	1.75 (1.62 to 1.90)*
	Adjusted	1 (Reference)	1.24 (1.10 to 1.40)*	1.20 (1.09 to 1.31)*
	ORL	1065 (18.1)	584 (29.5)	791 (25.5)
Unadjusted	1 (Reference)	1.88 (1.68 to 2.12)*	1.55 (1.39 to 1.72)*	
Adjusted	1 (Reference)	1.20 (1.05 to 1.37)*	1.16 (1.03 to 1.30)*	
Critical 90-day complication	All surgeries	2974 (3.3)	1577 (6.8)	2374 (5.2)
	Unadjusted	1 (Reference)	2.15 (2.02 to 2.29)*	1.62 (1.53 to 1.71)*
	Adjusted	1 (Reference)	1.09 (1.02 to 1.17)*	1.21 (1.14 to 1.29)*

Values are n (%) unless otherwise indicated. \*Odds ratio (95% c.i.). Gastro, gastroenterological surgery; Ortho, orthopaedic surgery; GYN, gynaecological surgery; Plastics, plastic surgery; ORL, otorhinolaryngological surgery.

all specialties (OR 1.21 (95 per cent c.i. 1.13 to 1.29) in gastroenterological surgery, OR 1.36 (95 per cent c.i. 1.27 to 1.46) in orthopaedic surgery, OR 1.12 (95 per cent c.i. 1.03 to 1.21) in gynaecological surgery, OR 1.20 (95 per cent c.i. 1.09 to 1.31) in plastic surgery, and OR 1.16 (95 per cent c.i. 1.03 to 1.30) in otorhinolaryngological surgery). The unadjusted odds of critical complications in all surgeries were increased for both ex-smokers and current smokers, and this was also true after

adjustment, with OR 1.09 (95 per cent c.i. 1.02 to 1.17) for ex-smokers and OR 1.21 (95 per cent c.i. 1.14 to 1.29) for current smokers.

When assessing relative variable importance, ASA class was the most important complication predictor in a majority of the models (Table 3). Concerning overall complications, smoking status had mean Shapley values (absolute log odds scale) of 0.07 in all surgeries, 0.08 in gastroenterological surgery, 0.13 in orthopaedic

Table 3 Relative variable importance

Outcome	Specialty	Statistic	Variable					MPR <sup>2</sup>	NPR <sup>2</sup>
			SS	Age (years)	Sex (male)	ASA	CCI		
Any 90-day complication	All surgeries	Wald X	141.4	209.3	584.4	10 437.4	1425.2	0.12	0.20
		Wald X (%)	1.1	1.6	4.6	81.6	11.1		
		Shapley	0.07	0.12	0.14	0.79	0.24		
	Gastro	Wald X	34.2	4.9	15.2	1820.7	370.6	0.13	0.21
		Wald X (%)	1.5	0.20	0.70	81.1	16.5		
		Shapley	0.08	0.04	0.05	0.78	0.28		
	Ortho	Wald X	76.8	10.8	88.6	734.0	112.4	0.07	0.12
		Wald X (%)	7.5	1.1	8.7	71.8	11.0		
		Shapley	0.13	0.07	0.15	0.52	0.12		
	GYN	Wald X	9.0	92.0	17.1	462.6	102.8	0.04	0.06
		Wald X (%)	1.3	13.5	2.5	67.7	15.0		
		Shapley	0.04	0.21	0.01	0.38	0.13		
	Plastics	Wald X	21.1	19.5	171.9	753.6	31.3	0.12	0.20
		Wald X (%)	2.1	2.0	17.2	75.6	3.1		
		Shapley	0.10	0.12	0.26	0.78	0.12		
ORL	Wald X	10.1	0.60	26.8	153.5	498.8	0.11	0.17	
	Wald X (%)	1.5	0.10	3.9	22.3	72.3			
	Shapley	0.08	0.03	0.13	0.31	0.47			
Critical 90-day complication	All surgeries	Wald X	39.1	29.6	15.3	5386.1	538.4	0.19	0.22
		Wald X (%)	0.70	0.50	0.30	89.6	9.0		
		Shapley	0.09	0.11	0.05	1.2	0.29		

Gastro, gastroenterological surgery; Ortho, orthopaedic surgery; GYN, gynaecological surgery; Plastics, plastic surgery; ORL, otorhinolaryngological surgery; SS, smoking status; CCI, Charlson co-morbidity index; MPR<sup>2</sup>, McFadden's pseudo R<sup>2</sup>; NPR<sup>2</sup>, Nagelkerke's pseudo R<sup>2</sup>.

surgery, 0.04 in gynaecological surgery, 0.10 in plastic surgery, and 0.08 in otorhinolaryngological surgery. The corresponding Wald (X) % values were 1.1, 1.5, 7.5, 1.3, 2.1, and 1.5 respectively. In the critical complication model, smoking status had a mean Shapley value of 0.09 and a Wald (X) % value of 0.70. The overall complication model including all surgeries had a McFadden's pseudo R<sup>2</sup> of 0.12 and a Nagelkerke's pseudo R<sup>2</sup> of 0.20.

## Discussion

In this large database study, an MLA for smoking status classification based on EHR notes was successfully developed. Both former and current smoking significantly increased the risk of overall 90-day postoperative complications in all surgeries. Ex-smokers had a significantly increased overall 90-day complication risk in gastroenterological, plastic, and otorhinolaryngological surgery. For current smokers, this risk was increased in all secondary analysis specialties (gastroenterological, orthopaedic, gynaecological, plastic, and otorhinolaryngological surgery). The risk of 90-day critical complications in all types of surgery was increased in both former and current smokers.

Compared with the Michigan study, the current population had a somewhat larger proportion of current smokers (25 versus 29 per cent)<sup>2</sup>. In the general Finnish population in 2017, 13 per cent of men and 10 per cent of women aged 20 to 84 years were daily smokers<sup>22</sup>. It must be stressed that a large group of patients with unknown smoking status were excluded from the analyses, and the true proportion of current smokers in the sample may differ from this study's results. However, based on both this study and previous ones, the smoking problem among surgical patients is apparent. The smoking trend in this study was downward. This may reflect changes in attitudes towards smoking and effects of increased tobacco taxation.

EHR free-text notes are often heterogeneous and inexact, and distinguishing between current and ex-smokers as per this study's definition often proved impossible and made the

classification task complex. Examples of sentences classified into the unknown category are 'has smoked before', 'has smoked x pack-years', 'long smoking history', and 'stopped smoking recently'. Sentences such as 'stopped smoking 3/20', 'stopped smoking in the 60s', and 'smoked 1995–2011' were considered accurate enough to be classified into the ex-smoker category.

To reliably assess smoking status as a predictor of surgical complications, this study aimed to maximize the reliability of smoking status classifications at the expense of a higher number of samples for statistical analysis. In other words, the high proportion of true positives (precision) for the ex-smoker, current smoker, and never-smoker classes was emphasized, while achieving a high recall for the unknown sentences. Therefore, sentences with unknown smoking status were also included in the training of the MLA and thresholds were set for the probabilities that were required for a sentence to be classified as an ex-smoker, current smoker, or never-smoker. Because of this and the strict criteria set for classifying ex-smokers, direct comparison of this study's results with those of other studies utilizing MLA methods for smoking status classification is difficult.

In their Finnish study on the effect of smoking status on cancer patient mortality rate, Karlsson *et al.*<sup>23</sup> also used MLA methods for smoking status classification. Algorithms used were BERT and ULMFiT, which require more computing power than fastText. The BERT model performed the best, with an average precision of 88.2 per cent for the classes ex-smoker, smoker, and never-smoker, compared with the current study precision of 96.1 per cent. Karlsson *et al.*<sup>23</sup> reported performance metrics at the patient level while performing a performance assessment at the sentence level consisting of only unique sentences.

Hawn *et al.*<sup>24</sup> were the first ones to investigate smoking-related postoperative complications after major surgery in a large study population. Similar to this study, the main finding was increased rates of complications in both current and ex-smokers compared with never-smokers in adjusted analyses. In contrast,

the authors reported complication subgroups and stratified the population by pack-year, finding an exposure greater than 20 pack-years to significantly increase the complication risk. However, non-cardiac and emergency cases were excluded, and multivariable analyses did not adjust for sex or co-morbidities.

Turan *et al.*<sup>25</sup> chose to propensity-match current smokers with never-smoker controls, utilizing data on major surgeries from a national quality improvement programme. Current smokers had an increased likelihood of 30-day mortality rate and major complications, which is consistent with the results of this study. In addition, a linear relationship between the odds of any major morbidity rate and the amount of smoking was found in a smaller subset of matched current smokers. In contrast, an ex-smoker group was not included, and patients with severe preoperative disease were excluded.

In a comprehensive study in 2013, Musallam *et al.*<sup>26</sup> investigated the effects of current and former smoking on postoperative mortality rate and vascular and respiratory events after major surgery. Only current smokers had increased odds of mortality rate, whereas both current and ex-smokers had increased odds of arterial and respiratory events. The odds of arterial and respiratory events increased with pack-year exposure.

Schmid *et al.*<sup>27</sup> investigated the effects of smoking on postoperative complications in 16 major cardiovascular, orthopaedic, and oncological surgical procedures. Current smokers had increased odds of overall, pulmonary, wound, and septic complications after most cardiovascular and oncological procedures compared with never-smokers. Ex-smokers also had increased odds of such complications, but to a lesser extent.

In orthopaedic surgery, smoking-related problems include non-union of bones in fracture surgery<sup>28</sup>, surgical site infection in spine surgery<sup>29</sup>, and higher mortality rate and analgesia usage in hip and knee arthroplasty<sup>30</sup>. In gastroenterological surgery, smoking has been found to increase anastomotic leakage after colorectal surgery<sup>31–33</sup>, infections during pancreatic resection<sup>34</sup>, and the risk of incisional hernia after laparotomy<sup>35</sup>. In gynaecological surgery, the risks of smoking have not been studied extensively. In observational study settings, smoking has been found to increase the risk of incisional hernia after open gynaecological surgery<sup>36</sup> and severe complications after ovarian cancer surgery<sup>37</sup>. In plastic surgery, smoking has been recognized as a risk factor for wound complications, including infections, seroma, flap haematoma, flap necrosis, and dehiscence<sup>38</sup>. In otorhinolaryngological surgery, both former and current smoking increase the risk of wound complications after tumour resection involving total laryngectomy<sup>39</sup>. In otological procedures, smoking increases the risk of wound infections, wound dehiscence, and hospital readmission<sup>40</sup>. Complication subgroups were not analysed in this study, but the increased odds of overall complications in current smokers undergoing procedures of the aforementioned specialties are in line with the literature.

Based on the Shapley value-based global importance scores of this study's relative variable importance models, smoking status is associated with postoperative complications to roughly the same extent as age, with some variation between the models. The strongest association between smoking status and complications was found in orthopaedic surgery. These findings were the same when calculating partial chi-squared statistics for the variables. This is noteworthy considering that smoking was the only influenceable variable in the models.

Ex-smokers were defined as having had quit smoking at least 6 months previously, and current smokers had higher odds of complications than never-smokers. However, the initiation of

smoking cessation as early as 6 months before surgery is not realistic in most cases. This raises questions about the optimal timing and practices of cessation. Studies have shown that a preoperative smoking cessation interval of at least 4 weeks reduces complication rates. Each additional week increases the magnitude of effect by 19 per cent<sup>41</sup>. Short-term smoking cessation of less than 4 weeks does not seem to increase or lower the risks compared with current smoking<sup>42,43</sup>. However, no evidence indicates that patients cannot be advised to quit smoking at any time before surgery<sup>43</sup>. When it comes to smoking cessation support, evidence shows that intensive interventions should be initiated at least 4 weeks before surgery. The combination of behavioural support and pharmacotherapy has the largest effect on complication risk reduction<sup>44</sup>.

This study has some important strengths. First, the sample was large, and included a wide range of both inpatient and outpatient surgical procedures of variable invasiveness, providing valuable information on the effects of smoking in all types of surgery. The findings also demonstrate smoking-associated risks in gynaecological surgery, a sparsely studied topic. Second, a 90-day postoperative follow-up interval enables capturing of both early and late postoperative complications compared with a 30-day interval. Third, the utilization of an MLA was a novel approach to smoking status classification, and the results show that it is feasible in a population with EHRs in Finnish. Considering the challenge of distinguishing between current and ex-smokers, the predictive model performed well and required comparatively little computational power.

The study also has limitations. The data lacked information on pack-years and the number of cigarettes smoked, and the study findings rely on self-reported smoking status. The MLA did not consider tobacco products other than cigarettes. Due to the 6-month limit and the sometimes inexact EHR notes, the distinction between current smokers and ex-smokers sometimes proved challenging. Aiming for a classification as exact as possible led to the exclusion of a large group of patients with unknown smoking status from the analyses, which may have caused selection bias. In addition, the results of the MLA are not directly applicable to other languages than Finnish. With these limitations in mind, the study findings should be interpreted with some degree of caution. The inadequate smoking status recording in the EHRs in this study is certainly an important finding. At present, smoking status is not recorded in a systematic, structural, and extractable way in the HUS area. A future quality register could be a way to overcome this issue.

Based on the results of this study, MLAs appear feasible for smoking status classification in large data sets. Current and former smoking associate with complications in all types of inpatient and outpatient surgery of various invasiveness, and also with critical complications. The results suggest that the risks of surgical complications associated with smoking and increasing age are approximately equal. However, of these two, smoking is the only modifiable one. In the future, intervention studies on efficient preoperative smoking cessation models are needed in the surgical field as a whole, and in orthopaedic surgery in particular.

## Funding

J.A.A. and O.H. are employed by Medaffcon Oy. The work of H.L.G. was supported by the Research Foundation of the Pulmonary Diseases, the Swedish Cultural Foundation in Finland and the Suomen Tuberkuloosin Vastustamisyhdistyksen säätiö. The

fundings were not involved in study design, data collection, data analysis, manuscript preparation, and publication decisions. Open access funded by Helsinki University Library.

## Acknowledgements

H.L.G., J.A.A., and O.H. had full access to the data and take responsibility for data integrity and accuracy of the data analysis.

Contributors: the authors thank Dr Jaana Ahlmaa for critical evaluation of the manuscript prior to submission and Dr Jonas Nordström for his help with the clinical note sentence annotation.

## Author contributions

Helene L. Gräsbeck (Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing), Aleksi R. P. Reito (Formal analysis, Investigation, Methodology, Visualization), Heikki J. Ekroos (Conceptualization, Funding acquisition, Methodology, Project administration, Supervision), Juhani A. Aakko (Formal analysis, Methodology, Software, Supervision, Validation), Olivia Hölsä (Formal analysis, Methodology, Software, Validation), and Tuula M. Vasankari (Conceptualization, Methodology, Project administration, Supervision)

## Disclosure

The authors declare no conflicts of interest.

## Supplementary material

Supplementary material is available at BJS Open online.

## Data availability

The data sets generated and analysed during the study are not publicly available, but are available from the corresponding author on request.

## References

1. WHO Report on the Global Tobacco Epidemic 2021: Addressing New and Emerging Products. Geneva: WHO, 2021
2. Howard R, Singh K, Englesbe M. Prevalence and trends in smoking among surgical patients in Michigan, 2012–2019. *JAMA Netw Open* 2021;**4**:e210553
3. Benowitz NL, Gourlay SG, Francisco S. Cardiovascular toxicity of nicotine: implications for nicotine replacement therapy. *J Am Coll Cardiol* 1997;**29**:1422–1453
4. Silverstein P. Smoking and wound healing. *Am J Med* 1992;**93**: 22S–24S
5. Myles P, Iacono G, Hunt J, Fletcher H, Morris J, McIlroy D et al. Risk of respiratory complications and wound infection in patients undergoing ambulatory surgery: smokers versus nonsmokers. *Anesthesiology* 2002;**97**:842–847
6. Grønkjær M, Eliassen M, Skov-Ettrup LS, Tolstrup J, Christiansen A, Mikkelsen S et al. Preoperative smoking status and postoperative complications: a systematic review and meta-analysis. *Ann Surg* 2014;**259**:52–71
7. Nolan MB, Warner DO. Safety and efficacy of nicotine replacement therapy in the perioperative period: a narrative review. *Mayo Clin Proc* 2015;**90**:1553–1561
8. Collobert R, Weston J, Com J, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. *J Mach Learn Res* 2011;**12**:2493–2537
9. Patel J, Siddiqui Z, Krishnan A, Thyvalikakath TP. Leveraging electronic dental record data to classify patients based on their smoking intensity. *Methods Inf Med* 2019;**57**:253–260
10. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson E et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inf Decis Mak* 2019;**19**:1
11. Palmer E, Hassanpour S, Higgins J, Doherty J, Onega T. Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC Med Inf Decis Mak* 2019;**19**:141
12. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg* 2004;**240**:205–213
13. Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* 2011;**173**:676–682
14. Gasparini A. Comorbidity: an R package for computing comorbidity scores. *J Open Source Softw* 2018;**3**:648
15. West R. Tobacco smoking: health impact, prevalence, correlates and interventions. *Psychol Health* 2017;**32**:1018–1036
16. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. Proceedings of the Fifteenth Conference of the European Chapter of the Association for Computational Linguistics. 2017;**2**:427–431
17. Hölsä O. Machine learning-based classification of clinical notes to extract smoking status from electronic health records. Thesis, School of Science, Aalto University, 2022
18. Strumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst* 2014;**41**:647–665
19. Greenwell B. *fastshap: Fast Approximate Shapley Values*. <https://github.com/bgreenwell/fastshap> (accessed 24 February 2022)
20. van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009
21. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2020. <https://www.R-project.org/> (accessed 24 February 2022)
22. *Smoking in Finland—Alcohol, Tobacco and Addictions—THL*. <https://thl.fi/en/web/alcohol-tobacco-and-addictions/tobacco/smoking-in-finland> (accessed 14 May 2019)
23. Karlsson A, Ellonen A, Irjala H, Väliäho V, Mattila K, Nissi L et al. Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit. *ESMO Open* 2021;**6**:100175
24. Hawn MT, Houston TK, Campagna EJ, Graham L, Singh J, Bishop M et al. The attributable risk of smoking on surgical complications. *Ann Surg* 2011;**254**:914–920
25. Turan A, Mascha EJ, Roberman D, Turner P, You J, Kurz A et al. Smoking and perioperative outcomes. *Anesthesiology* 2011;**114**: 837–846
26. Musallam KM, Rosendaal FR, Zaatari G, Soweid A, Hoballah JJ, Sfeir PM. Smoking and the risk of mortality and vascular and respiratory events in patients undergoing major surgery. *JAMA Surg* 2013;**148**:755–762
27. Schmid M, Sood A, Campbell L, Kapoor V, Dalela D, Klett D et al. Impact of smoking on perioperative outcomes after major surgery. *Am J Surg* 2015;**210**:221–229.e6
28. Scolaro JA, Schenker ML, Yannascoli S, Baldwin K, Mehta S, Ahn J. Cigarette smoking increases complications following fracture: a systematic review. *J Bone Joint Surg Am* 2014;**96**:674–681

29. Xing D, Ma JX, Ma XL, Song D, Wang J, Chen Y et al. A methodological, systematic review of evidence-based independent risk factors for surgical site infections after spinal surgery. *Eur Spine J* 2013;**22**:605–615
30. Matharu GS, Mouchti S, Twigg S, Delmestri A, Murray D, Judge A et al. The effect of smoking on outcomes following primary total hip and knee arthroplasty: a population-based cohort study of 117,024 patients. *Acta Orthop* 2019;**90**:559–567
31. McDermott FD, Heeney A, Kelly ME, Steele RJ, Carlson GL, Winter DC. Systematic review of preoperative, intraoperative and postoperative risk factors for colorectal anastomotic leaks. *Br J Surg* 2015;**102**:462–479
32. Baucom RB, Poulouse BK, Herline AJ, Muldoon RL, Cone MM, Geiger TM. Smoking as dominant risk factor for anastomotic leak after left colon resection. *Am J Surg* 2015;**210**:1–5
33. Parthasarathy M, Greensmith M, Bowers D, Groot-Wassink T. Risk factors for anastomotic leakage after colorectal resection: a retrospective analysis of 17 518 patients. *Colorectal Dis* 2017;**19**:288–298
34. Okano K, Hirao T, Unno M, Fujii T, Yoshitomi H, Suzuki S et al. Postoperative infectious complications after pancreatic resection. *Br J Surg* 2015;**102**:1551–1560
35. Sørensen LT, Hemmingsen UB, Kirkeby LT, Kallehave F, Jørgensen LN. Smoking is a risk factor for incisional hernia. *Arch Surg* 2005;**140**:119–123
36. Bewö K, Österberg J, Löfgren M, Sandblom G. Incisional hernias following open gynecological surgery: a population-based study. *Arch Gynecol Obstet* 2019;**299**:1313–1319
37. Egger EK, Kohls N, Stope MB, Condic M, Keyver-Paik M, Könsgen D et al. Risk factors for severe complications in ovarian cancer surgery. *In Vivo* 2020;**34**:3361–3365
38. Theocharidis V, Katsaros I, Sgouromallis E, Serifis N, Boikou V, Tasigiorgos S et al. Current evidence on the role of smoking in plastic surgery elective procedures: a systematic review and meta-analysis. *J Plast Reconstr Aesthet Surg* 2018;**71**:624–636
39. Goepfert RP, Hutcheson KA, Lewin JS, Desai NG, Zafereo ME, Hessel AC et al. Complications, hospital length of stay, and readmission after total laryngectomy. *Cancer* 2017;**123**:1760–1767
40. Kay-Rivest E, Mascarella M, Sewitch MJ, Cloutier F, Mijovic T. Association between smoking and 30-day outcomes in otologic surgery. *Otolaryngol Head Neck Surg* 2020;**162**:108–113
41. Mills E, Eyawo O, Lockhart I, Kelly S, Wu P, Ebbert JO. Smoking cessation reduces postoperative complications: a systematic review and meta-analysis. *Am J Med* 2011;**124**:144–154.e8
42. Wong J, Lam DP, Abrishami A, Chan MTV, Chung F. Short-term preoperative smoking cessation and postoperative complications: a systematic review and meta-analysis. *Can J Anaesth* 2012;**59**:268–279
43. Myers K, Hajek P, Hinds C, McRobbie H. Stopping smoking shortly before surgery and postoperative complications: a systematic review and meta-analysis. *Arch Intern Med* 2011;**171**:983–989
44. Thomsen T, Villebro N, Møller AM. Interventions for preoperative smoking cessation. *Cochrane Database Syst Rev* 2014;**3**:CD002294