Journal of Clinical Epidemiology 118 (2020) 100−106

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Data mining information from electronic health records produced high yield and accuracy for current smoking status

T. Katrien J. Groenhof[a,*], Laurien R. Koers[a], Enja Blasse[b], Mark de Groot[b], Diederick E. Grobbee[a], Michiel L. Bots[a], Folkert W. Asselbergs[c,d,e], A. Titia Lely[f], Saskia Haitjema[b], On behalf of the UPOD, and UCC-CVRM Study Groups

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[b]*Laboratory of Clinical Chemistry and Hematology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[c]*Health Data Research UK, Institute of Health Informatics, University College London, London, UK*
[d]*Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[e]*Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, UK*
[f]*Department of Obstetrics, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*

Accepted 6 November 2019; Published online 12 November 2019

Abstract

Objectives: Researchers are increasingly using routine clinical data for care evaluations and feedback to patients and clinicians. The quality of these evaluations depends on the quality and completeness of the input data.

Study Design and Setting: We assessed the performance of an electronic health record (EHR)-based data mining algorithm, using the example of the smoking status in a cardiovascular population. As a reference standard, we used the questionnaire from the Utrecht Cardiovascular Cohort (UCC). To assess diagnostic accuracy, we calculated sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV).

Results: We analyzed 1,661 patients included in the UCC to January 18, 2019. Of those, 14% (n = 238) had missing information on smoking status in the UCC questionnaire. Data mining provided information on smoking status in 99% of the 1,661 participants. Diagnostic accuracy for current smoking was sensitivity 88%, specificity 92%, NPV 98%, and PPV 63%. From false positives, 85% reported they had quit smoking at the time of the UCC.

Conclusion: Data mining showed great potential in retrieving information on smoking (a near complete yield). Its diagnostic performance is good for negative smoking statuses. The implications of misclassification with data mining are dependent on the application of the data.

Keywords: Data mining; Electronic health records; Routine clinical data; Learning healthcare system; Data quality; Text mining

1. Introduction

Information recorded in the electronic health records (EHRs) has the possibility of revolutionizing our health care system into a ''Learning Healthcare System'' [1], in which routine clinical care and science are aligned via a constant cycle of data assembly, data analysis, interpretation, feedback, and change implementation [2]. EHRs contain routinely collected care information on symptoms, diagnosis, laboratory tests, other diagnostic tests, and treatments and are therefore a potential source of data for epidemiologic studies, pragmatic trials, drug safety evaluations, and health care organization evaluations [3].

For all (scientific) evaluations, the validity of results depends on the quality of input data. Current scientific evaluations mostly depend on randomized controlled trials and cohort studies. But these studies are affected by selection and nonresponse, sparking interest in the use of routine care

Declaration of interest: The UCC is primarily financed by the UMC Utrecht. A grant from the Netherlands Organisation for Health Research and Development (#8480-34001) was obtained to develop feedback procedures. UCC website: www.umuctrecht.nl/ucc (in Dutch). Contact information of UCC: ucc@umcutrecht.nl.

* Corresponding author. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands. Tel.: +31-88-75-69308; fax: +31-88-75-68099.
E-mail address: t.k.j.groenhof@umcutrecht.nl (T.K.J. Groenhof).

https://doi.org/10.1016/j.jclinepi.2019.11.006

What is new?

Key findings

- Data mining showed great potential in retrieving information on smoking from the electronic health record (EHR). Its diagnostic performance is good for negative smoking statuses.

What this adds this to what is known?

- Via data mining we can successfully extract information from both structured and unstructured fields in the EHR for scientific evaluations. Data quality evaluation, comparing the EHR information to a reference standard, should be part of the mining process.

What is the implication and what should change now?

- If EHR-based data mining algorithms are used to retrieve information for care or scientific purposes, the effect of time and clinical practice on the outcome, and the implications of misclassification need to be taken into account.

data for research purposes [4]. Routinely collected clinical care data are not restricted by selection or nonresponse and thus reflects the real-world situation better [4]. Yet, EHR data consist of both structured and unstructured data: the clinical notes. Clinical notes are case-specific notes that capture nuances and clinical reasoning. Apart from large heterogeneity among clinicians, clinical notes are prone to spelling errors, abbreviations, inconsistencies, and idiosyncrasies in a complex context. Also, clinical notes can be subject to missing, which is usually not at random (confounding by indication). Careful evaluation of EHR data quality and applicability needs to be considered when used for (scientific and clinical) evaluations [5].

In the past years, many data mining algorithms have been developed to extract data from the EHR [6]. The information yield is large, but data quality is underreported. As a proof of concept, we evaluated an EHR-based data mining algorithm, applied to the case of smoking status (never, ever, current). Smoking status is of large clinical importance for it may be used to identify those at elevated risk of disease in risk prediction algorithms, for example cardiovascular disease, the results of which has consequences for clinical decision-making [7]. Furthermore, smoking, as important causal factor in many noncommunicable diseases, is of great importance for etiologic and prognostic research questions as a confounder, modifier, or predictor. Smoking status can potentially be registered in several locations within the EHR: structured questionnaires such as intoxication boxes (predefined answers to the question "smoking?"), unstructured questionnaires ("smoking?" with a free text response field), and unstructured free text including clinical notes and (discharge) letters. Especially with a characteristic such as smoking that is subjective, might change over time, and might feel stigmatizing, discrepancies in reporting can easily occur. We wondered if we could apply data mining to retrieve the smoking status from our EHR and provide insight into the quality of the information.

2. Methods

2.1. Study design and population

We performed a cross-sectional analysis, using data from the Utrecht Cardiovascular Cohort (UCC) and the EHR from the University Medical Center Utrecht. In short, the UCC is a prospective cohort study targeted to uniform assessment and registration of the guideline-based cardiovascular risk profile in all patients presenting with a (risk factor for) cardiovascular disease within routine care [8]. The UCC has been approved by the Institutional Review Board Biobank of the UMC Utrecht, and all data are handled according to privacy regulations [2]. We used data from patients included in the cohort up to January 28, 2019, who had provided a written informed consent.

2.2. Data collection and definitions

The UCC smoking status was retrieved via a questionnaire that was either filled in at home via the patient portal or on paper and registered within the EHR. Two questions were asked on smoking: *"Do you currently smoke?"* and *"Did you smoke in the past?"* which resulted in three smoking status categories: current smoker, past smoker, and never smoker. To assess diagnostic accuracy, we dichotomized this into current smoker and nonsmoker, the latter including both past and never smokers.

The data mining algorithm can be best described as a decision rule model. First, we mined information on smoking status in UCC patients' EHR, with exclusion from the UCC questionnaire (also registered within the EHR). This information is captured in structured fields, including specialism-specific questionnaires and the UMC Utrecht-wide intoxications field, or in–unstructured-free text, including clinical notes, correspondence from and to colleagues. Second, all free text and structured text fields from 365 days before to 7 days after the UCC were selected, with the exception from letters after discharge, which were searched within the week after the UCC (Fig. 1). Third, retrieved information was categorized into current smokers, past smokers, never smokers, and unknown. Structured text statuses were directly categorized. Free-text fragments were used to build text constructs: first a keyword word (smoking, smoked, smoker, …) was selected, then surrounding sentence fragments were assessed for interpretation (±, sometimes, quit, …) and the statuses were

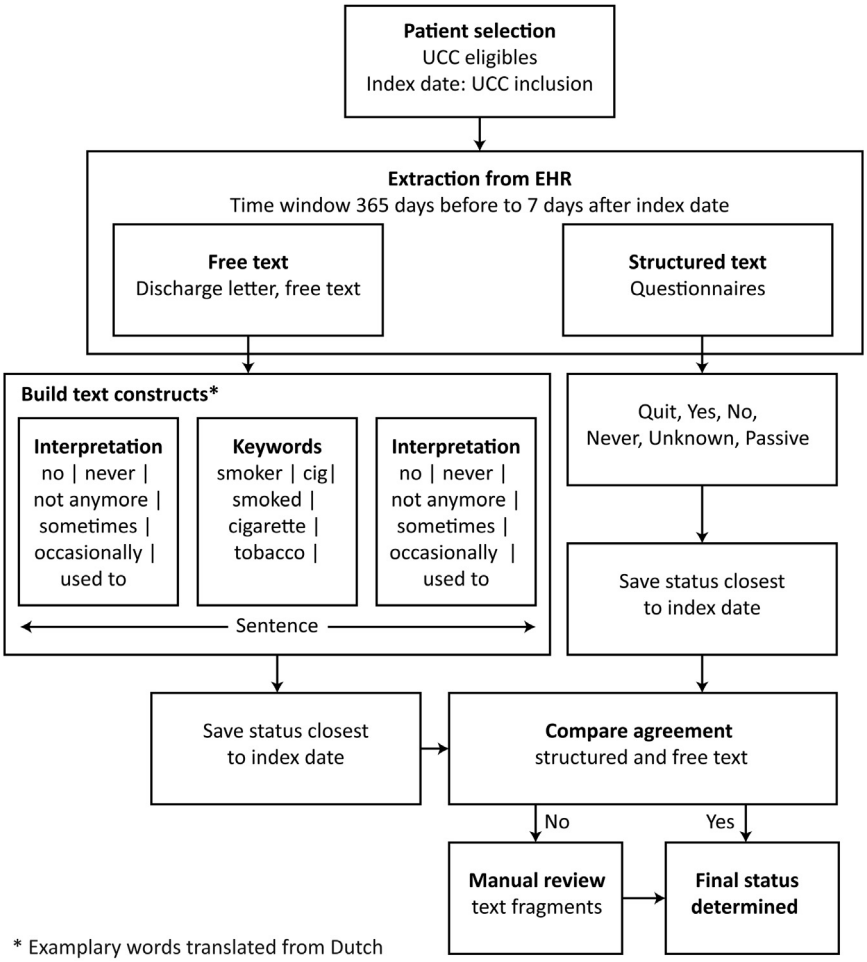


Fig. 1. Data mining algorithm.

finalized. Fourth, the status closest to the inclusion date was included in the analyses. Lastly, statuses from free and structured text were compared for agreement. This also included a sanity check: patients who reported current smoking before a ''never smoked'' status were redefined as ''quit smoking'' in the latter status. If the algorithm did not achieve agreement or was inconclusive, a manual check of the EHR text fragments was conducted to finalize the status. To assess diagnostic accuracy, we recoded a new smoking variable similar to the categorization in the UCC.

Patient characteristics were derived from the UCC questionnaire, and anthropometric, blood pressure, and laboratory measurements were collected during the routine hospital visit. In the UCC, we collected information on sex, age, body mass index (BMI), origin, level of education, specialism of inclusion, established cardiovascular disease, and risk factors for cardiovascular disease. Established cardiovascular disease was defined as a history of a coronary heart disease, cerebrovascular disease, peripheral artery disease, and/or aneurysm of the abdominal aorta. A history of coronary heart disease was defined as myocardial infarction or coronary revascularization procedures. A history of cerebrovascular disease was defined as ischemic stroke, transient ischemic attack, cerebral hemorrhage, and/or carotid stenosis. A history of peripheral artery disease was defined as intermittent claudication with a vascular cause and/or peripheral artery revascularization procedures. A history of an abdominal aortic aneurysm was defined as an abdominal aortic aneurysm requiring surgery. Risk factors for cardiovascular disease were hypertension, defined as a positive history of hypertension, prescription of blood pressure lowering medication, and/or a blood pressure higher than RR140/90 mmHg; and diabetes mellitus, defined as a positive history of diabetes mellitus, prescription of blood glucose lowering medication, and/or a HbA1c above 48 mmol/mol.

2.3. Analyses

The yield of data mining was defined as the percentage of UCC patients for whom a smoking status was retrieved.

Diagnostic accuracy was assessed comparing the data mining status with the reference test: the UCC status based on questionnaire responses. We calculated sensitivity,

Table 1. Patient characteristics

Patients, *n*	Total *n* = 1,661
Women, *n* (%)	769 (46)
Age (years), median (IQR)	60 (17)
Dutch origin, *n* (%)	1,205 (73)
Level of education, *n* (%)	
Lower	499 (31)
Intermediate	355 (22)
High	536 (33)
Missing	271 (16)
Specialism of inclusion, *n* (%)	
Cardiology	665
Cardiothoracic surgery	1
Diabetes	110
Geriatrics	325
Infectious diseases	14
Nephrology	94
Vascular medicine	134
Neurology	257
Vascular surgery	1
OBGYN	60
Manifest cardiovascular disease, *n* (%)	914 (55)
Risk factors, *n* (%)	
Hypertension	701 (42)
Diabetes	320 (19)
Smoking	
Current	189 (11)
Quit	674 (41)
Never	560 (34)
Missing	238 (14)

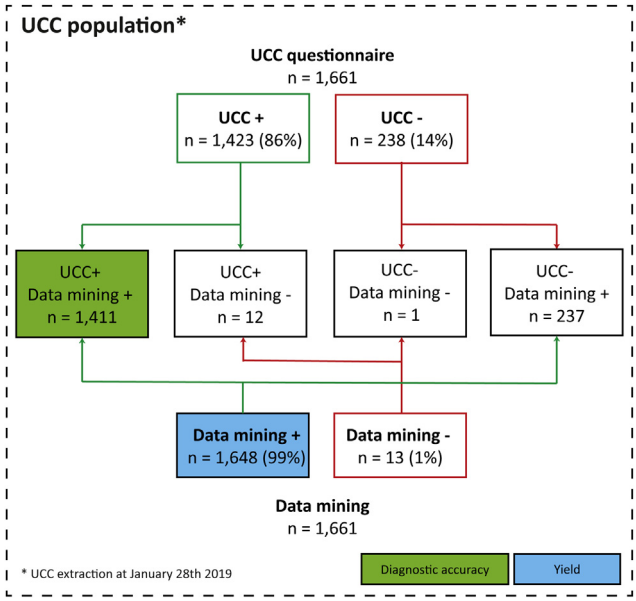
specificity, negative predictive value (NPV), and positive predictive value (PPV).

Additionally, we assessed factors associated with misclassification of the data mining outcome, defined as either a false-positive or false-negative status. We performed a logistic regression analysis with misclassification as the outcome and specialism of the UCC inclusion, difference in days between UCC and data mining status, number of statuses found, different sources where statuses were found, and type of source where the final status was derived from as potential factors. For this analysis, we excluded specialisms with less than five inclusions, and the departments of obstetrics and gynecology were combined.

All analyses were performed in R Statistical Software version 4.3, Foundation for Statistical Computing, Vienna, Austria.

3. Results

For these analyses, we used data from UCC patients included in the cohort between January 2016 and January 28, 2019 (n = 1,661; Table 1). The population was mostly



Legend. UCC: Utrecht Cardiovascular Cohort, +: smoking status known, -: status unknown

Fig. 2. Patient selection and availability of smoking status from UCC and data mining.

of Dutch native origin (73%), with both men (54%) and women (46%). The median age was 60 years (interquartile range [IQR] 17). Most patients were included in the departments of cardiology (n = 665, 40%) and geriatrics (n = 325, 20%). Over half of patients (n = 914, 55%) had a history of a cardiovascular event, 42% suffered from hypertension and 19% from diabetes.

Within the UCC, smoking status was reported by 1,423 patients (85%, Fig. 2). From data mining, smoking status was available for 1,648 patients (99%, Fig. 2). For each patient, we found a median of 1.96 (range 1:15) number of smoking statuses in the EHR.

Information on smoking status (Table 2) was retrieved a median of 1.96 times per patient (minimum 1, maximum 15 times), and the sources were structured questionnaires (35%), letters (29%), and nonstructured questionnaires (36%). The median difference in days between mined and UCC smoking statuses was 0 days (minimum 0 days, maximum 360 days). Twelve patients did have a UCC status but no data mining status, and one patient did not have a UCC nor a data mining status: these patients came from a neighboring hospital, underwent a percutaneous transluminal coronary angioplasty (PTCA) in our center, and were discharged to that neighboring hospital directly after this procedure. The correspondence on the PTCA did only feature information about the procedure. This left us with 1,411 patients with both UCC and data mining smoking statuses (Fig. 2).

Diagnostic accuracy was assessed in 1,411 patients with both UCC and data mining smoking statuses (Table 3). The prevalence of current smoking was 11%. If we would include past smokers as smokers, the prevalence would increase to 60%. Sensitivity was 88%, specificity 92%, NPV

Table 2. Mining characteristics

Mining characteristics	Total n = 1,611
Number of smoking status retrieved, median (min-max)	1.96 (1.0–15.0)
Number of status sources per patient, n (%)	
1	874 (53)
2	602 (36)
3	185 (11)
Sources of final status, n (%)	
Structured questionnaire	590 (35)
Letter	475 (29)
Unstructured questionnaire	596 (36)
Difference in days between mined and UCC status, median (min-max)	0 (0; 360)

98%, and PPV was 63%. From 97 false positives, 82 (85%) of patients explicitly reported they had quit smoking in the UCC questionnaire. If we would include these 82 patients as current smokers, the PPV for current smoking would increase to 94% (Table 4). If we would exclude patients with a manually reviewed status, diagnostic performance would remain similar, only precision would decrease due to the decrease in sample size.

For analysis on factors associated with misclassification, we excluded departments with less than five patients included in the UCC: both vascular and cardiothoracic surgery departments were excluded because they included one patient. This left us with 1,409 patients for this analysis (Table 5). We found higher odds for misclassification in patients with multiple smoking statuses from multiple sources. Compared with information retrieved from structured questionnaires, smoking status was misclassified less often in letters (OR 0.45 [95% CI] 0.26; 0.77). Compared with the department of vascular medicine, all departments, except the department of infectious diseases, had lower odds for misclassification, albeit not all reached statistical significance. The difference in days between data mining, and UCC status was not associated with misclassification.

4. Discussion

We assessed the performance of an EHR-based data mining algorithm to detect the smoking status in a cardiovascular population. Data mining showed great potential in retrieving information on smoking (a near complete yield). Its diagnostic performance is good for a nonsmoking status. The implications of misclassification with data mining depends on the application of the data.

Many data mining algorithms have been developed and published over the past years [6]. Four studies specifically described smoking status mining. The yield in these studies was between 68 and 94% [9–13] and reported similarities between reference standard and data mining algorithm varied from kappa 0.5 to 0.98 [12,13]. This variation might be explained by differences in information sources, type of mining (codes, free text, machine learning), and the study population. Two studies from the United Kingdom based their algorithms on *"read codes,"* a coded thesaurus of clinical terms that has been issued by the National Health Service (NHS) in 1985 and reported a yield of 94% [11,12]. Wu et al. used only free text mining, applied to a mental health register and reported a yield of 68% [9]. Patel et al. mined free text from electronic dental records using three machine learning classifiers and reported a yield of 76% [13]. Our data mining algorithm was based on both structured fields and free-text fragments, scoping the entire EHR without restrictions to specific departments. Also, we developed the algorithm together with clinicians, knowing exactly where and how this kind of data is registered. Involvement of clinicians in the development of decision tools and data analytic developments has proven to result in better performing tools before [14]. Furthermore, our algorithm was optimized with data derived from a cardiovascular cohort within our center (n = 1,200). Lastly, smoking is an important risk factor for cardiovascular disease; this should have been documented in the EHR for every cardiovascular patient, explaining our high yield.

The prevalence of current smoking was lower in our study (11%) compared with other cardiovascular populations (16%–33%) [15,16]. This might be explained by the decreasing trend in the prevalence of cardiovascular risk factors: current smoking was prevalent in 48% of a cardiovascular population between 1996 and 1998 and decreased over time to 25% in 2011–2014. Our cohort dates from 2016 to 2018 and follows this decrease in smoking prevalence. Past smoking in our cohort was 48%, which is in line with the past-smoking prevalence in the SMART cohort, another cardiovascular cohort from the UMC Utrecht

Table 3. 2 × 2 contingency table on current smoking status

	UCC	
Smoking status	**Current smoker**	**Nonsmoker**
Data mining		
Current smoker	165	97
Nonsmoker	22	1125

Sensitivity: 88% (95% CI 83%–92%); specificity: 92% (95% CI 90%–94%); positive predictive value (PPV: 63% [95% CI 58%–67]; negative predictive value (NPV 98% [95% CI 97%–98%].

Table 4. 3 × 3 contingency table on smoking status categories

	UCC		
Smoking status	**Current smoker**	**Quit smoking**	**Never smoked**
Data mining			
Current smoker	165	82	15
Quit smoking	12	394	33
Never smoked	10	191	507

82/97 = 85% of patients smoked in data mining data but quit by the time of the UCC questionnaire.

Table 5. Factors associated with misclassification of the smoking status[a]

Characteristics	OR (95% CI) $n = 1,409$
Number of locations where smoking information was retrieved	1.07 (0.92; 1.45)
Number of different sources	1.59 (1.09; 2.73)
Location of final status	
Structured questionnaire	Reference
Letter	0.45 (0.26; 0.77)
Unstructured questionnaire	0.75 (0.47; 1.21)
Specialism	
Vascular medicine	Reference
Cardiology	0.86 (0.46; 1.71)
Diabetes care	0.61 (0.24; 1.48)
Geriatrics	0.23 (0.09; 0.54)
Infectious diseases	1.54 (0.37; 2.35)
Nephrology	0.95 (0.37; 2.35)
Neurology	0.76 (0.32; 1.88)
OBGYN	0.62 (0.13; 2.09)
Difference in days between UCC and data mining status date	0.99 (0.991; 0.999)

Abbreviations: OR, odds ratio; 95% CI, 95% confidence interval; OBGYN, departments of obstetrics and gynecology.

An OR above 1 means higher odds of misclassification.

[a] Input variables: specialism of UCC inclusion, difference in days between UCC and data mining status, number of status retrieved, number of different sources mining statuses were retrieved from, different sources of the final status.

(50%) [17]. Compared with the sensitivity, specificity, and NPV, PPV we observed was much lower. This might be explained by some underreporting of smoking in the UCC questionnaire, for this is self-reported [18]. More importantly, some patients smoked according to the data mining status but quit at the time they filled in the UCC questionnaire. Possibly, this is due to time and clinical interference. If we would consider these quit smokers to be current smokers in the UCC, the PPV would increase to 94%.

The implication of the PPV we found in our study, and diagnostic performance of data mining algorithms in general, depends on for which purpose the data are being used. The application of mined data can be either clinical decision-making or research. If a mined smoking status is used for clinical decision-making including prediction, we might misclassify some patients as smokers although they just have recently quit smoking. Because the contribution to risk from smoking does not disappear overnight after smoking cessation, we think this overestimation of risk is negligible. If you would use it for risk assessment, we would recommend to ask the patient and take time since cessation into consideration when calculating a risk score so the risk is not underestimated for recently quit smokers. Because the risk from smoking might not yet have decreased to zero, preventive measures directed toward other risk factors might still be indicated. If a mined smoking status is used for research, we can think of three scenarios. First, smoking is used as a confounder: misclassification is only an issue in positive tests (e.g., the current smokers), but because these contain mostly quit smokers who just recently quit, the risk from smoking did not decrease yet and overestimation seems negligible. Second, smoking is used as a modifier: differentiation between past and current smokers can be a problem. Selective re-evaluation of only positive smoking statuses would increase accuracy and still be less laborious than collecting information on smoking for all patients. Third, smoking is used as determinant for etiologic evaluations with the effect of smoking being acute or chronic: misclassification can be problematic if current smoking is used and structured re-evaluation of positive statuses is recommended, but misclassification of chronic exposure, that is, packyears, is less influenced by misclassification. Yet, reliable selection of individuals that do not smoke is reliable. Thus, if EHR-based data mining algorithms are used to retrieve information for care or scientific purposes, the effect of time and clinical practice on the outcome, and the implications of misclassification need to be taken into account.

Our study has strengths and limitations. We were able to include patients from all departments in our center treating CVD patients, which makes our result generalizable to a large group. The algorithm mined the entire EHR of our patients, without restrictions to department or type of text, included all information on smoking that was retrieved, interpreted them in correlation to the other information on smoking that was retrieved from one patient and then included the information on smoking that was documented in the EHR closest in time to our reference test. The downside of mining multiple statuses was that we also found conflicting statuses from multiple sources, which was associated with more misclassification. Furthermore, we used the UCC questionnaire as our reference standard. It would have been preferable to have an objective reference test for smoking via blood or urine. Such a test is unfortunately not available. EHR data quality should be of main concern, and methods to improve this should be addressed in future research.

From a data analytics perspective, we applied a fairly simple rule-based decision tree mining technique in this project. Many more (complex) (clinical) text mining programs and packages are being developed such as tidytext in R and word2vec for Python [19,20]. Because these kinds of techniques rely on standard text principles including negations, stemming, and word order, they might be more easily scalable to other settings compared with the presented rule-based model that was trained on our local data [21]. The Institute of Electrical and Electronics Engineers (IEEE, https://www.ieee.org/) and Association for Computing Machinery (ACM, https://www.acm.org/) also present many advanced papers on text mining in general, for example on sentiment-analysis [22] and hidden medication patterns in EHRs [23]. The main focus of these papers

is, however, on mathematics and formulas, and efforts must be made to help translate these complex formulas into actionable data analytics understandable for the clinician. To serve this purpose, we are working on pipeline for these more advanced clinical text mining methods that is expected to be submitted for publication soon.

In conclusion, data mining showed great potential in retrieving information on smoking (nearly complete yield). Its diagnostic performance is good for negative smoking statuses. The implications of misclassification with data mining is dependent on the application of the data.

Acknowledgments

Members of the UPOD study group: Wouter van Solinge, Imo Hoefer, Saskia Haitjema, Mark de Groot. Members of the Utrecht Cardiovascular Cohort-CardioVascular Risk Management (UCC- CVRM) Study group: F.W. Asselbergs, Department of Cardiology; G.J. de Borst, Department of Vascular Surgery; M.L. Bots (chair), Julius Center for Health Sciences and Primary Care; S. Dieleman, Division of Vital Functions (anesthesiology and intensive care); M.H. Emmelot, Department of Geriatrics; P.A. de Jong, Department of Radiology; A.T. Lely, Department of Obstetrics/Gynecology; I.E. Hoefer, Laboratory of Clinical Chemistry and Hematology; N.P. van der Kaaij, Department of Cardiothoracic Surgery; Y.M. Ruigrok, Department of Neurology; M.C. Verhaar, Department of Nephrology & Hypertension, F.L.J. Visseren, Department of Vascular Medicine, University Medical Center Utrecht and Utrecht University.

References

[1] Olsen L, Aisner D, McGinnis JM. The learning healthcare system: workshop summary. Roundtabel on evidence-based medicine. Washington, DC: The National Academies Press; 2007.

[2] Foley T, Fairmichael F. The potential of learning healthcare systems. 2015. Available at http://www.learninghealthcareproject.org/LHS_Report_2015.pdf. Accessed December 9, 2019.

[3] Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016; 23(5):1007−15.

[4] Budrionis A, Bellika JG. The Learning Healthcare System: where are we now? A systematic review. J Biomed Inform 2016;64:87−92.

[5] Afzal MZ. Text mining to support knowledge discovery from electronic health records. The Netherlands: Erasmus University Rotterdam; 2018.

[6] Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: a review. J Biomed Inform 2017;70:1−13.

[7] Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. Rev Esp Cardiol (Engl Ed) 2016;69(10):939.

[8] Asselbergs FW, Visseren FL, Bots ML, de Borst GJ, Buijsrogge MP, Dieleman JM, et al. Uniform data collection in routine clinical practice in cardiovascular patients for optimal care, quality control and research: the Utrecht Cardiovascular Cohort. Eur J Prev Cardiol 2017;24(8):840−7.

[9] Wu CY, Chang CK, Robson D, Jackson R, Chen SJ, Hayes RD, et al. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. PLoS One 2013;8:e74262.

[10] Mant J, Murphy M, Rose P, Vessey M. The accuracy of general practitioner records of smoking an alcohol use: comparison with patient questionnaires. J Public Health Med 2000;22(2):198−201.

[11] Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, White IR, et al. Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study. BMJ Open 2014;4:e004958.

[12] Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, Brophy ST, et al. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. BMC Med Inform Decis Mak 2017;17:2.

[13] Patel J, Siddiqui Z, Krishnan A, Thyvalikakath TP. Leveraging electronic dental record data to classify patients based on their smoking intensity. Methods Inf Med 2018;57:253−60.

[14] Bezemer T, de Groot MC, Blasse E, Ten Berg MJ, Kappen TH, Bredenoord AL, et al. A human(e) factor in clinical decision support systems. J Med Internet Res 2019;21(3):e11732.

[15] Kaasenbrood L, Bhatt DL, Dorresteijn JAN, Wilson PWF, D'Agostino RB, Massaro JM, et al. Estimated life expectancy without recurrent cardiovascular events in patients with vascular disease: the SMART-REACH model. J Am Heart Assoc 2018;7(16):e009217.

[16] Dorresteijn JA, Visseren FL, Wassink AM, Gondrie MJ, Steyerberg EW, Ridker PM, et al. Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. Heart 2013; 99:866−72.

[17] Berkelmans GFN, van der Graaf Y, Dorresteijn JAN, de Borst GJ, Cramer MJ, Kappelle LJ, et al. Decline in risk of recurrent cardiovascular events in the period 1996 to 2014 partly explained by better treatment of risk factors and less subclinical atherosclerosis. Int J Cardiol 2018;251:96−102.

[18] Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res 2009;11(1):12−24.

[19] Silge J, Robinson D, Fay C, Benoit K, Kanishka. Tidytext. 2019. juliasilge/tidytext: tidytext 0.2. Available at http://github.com/juliasilge/tidytext. Accessed August 24, 2019.

[20] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR2013. Available at https://arxiv.org/abs/1301.3781. Accessed August 24, 2019.

[21] García-Pedrajas N, de Haro-García A. Scaling up data mining algorithms: review and taxonomy. Prog Artif Intellig 2012;1(1):71−87.

[22] Rida-e-fatima S, Javed A, Banjar A, Irtaza A, Dawood H, Dawood H, et al. A multi-layer dual attention deep learning model with refined word embeddings for aspect-based sentiment analysis. IEEE Access 2019;7:114795−807.

[23] Huang H-Q, Shang X-P, Zhao H-M, Wu N, Li W-Z, Xu Y, et al. Discovering medication patterns for high-complexity drug-using diseases through electronic medical records. IEEE Access 2019;7:125280−99.