




Data mining to retrieve smoking status from electronic health records in general practice

Annemarijn R. de Boer ^{1,2,*}, Mark C.H. de Groot³, T. Katrien J. Groenhof¹, Sander van Doorn ¹, Ilonca Vaartjes^{1,2}, Michiel L. Bots ¹, and Saskia Haitjema³

¹Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, Utrecht 3584 CX, The Netherlands; ²Dutch Heart Foundation, The Hague, The Netherlands; and ³Central Diagnostic Laboratory, University Medical Center Utrecht, Utrecht, The Netherlands

Received 23 December 2021; revised 19 April 2022; online publish-ahead-of-print 20 May 2022

Aims

Optimize and assess the performance of an existing data mining algorithm for smoking status from hospital electronic health records (EHRs) in general practice EHRs.

Methods and results

We optimized an existing algorithm in a training set containing all clinical notes from 498 individuals (75 712 contact moments) from the Julius General Practitioners' Network (JGPN). Each moment was classified as either 'current smoker', 'former smoker', 'never smoker', or 'no information'. As a reference, we manually reviewed EHRs. Algorithm performance was assessed in an independent test set ($n = 494$, 78 129 moments) using precision, recall, and F1-score. Test set algorithm performance for 'current smoker' was precision 79.7%, recall 78.3%, and F1-score 0.79. For former smoker, it was precision 73.8%, recall 64.0%, and F1-score 0.69. For never smoker, it was precision 92.0%, recall 74.9%, and F1-score 0.83. On a patient level, performance for ever smoker (current and former smoker combined) was precision 87.9%, recall 94.7%, and F1-score 0.91. For never smoker, it was 98.0, 82.0, and 0.89%, respectively. We found a more narrative writing style in general practice than in hospital EHRs.

Conclusion

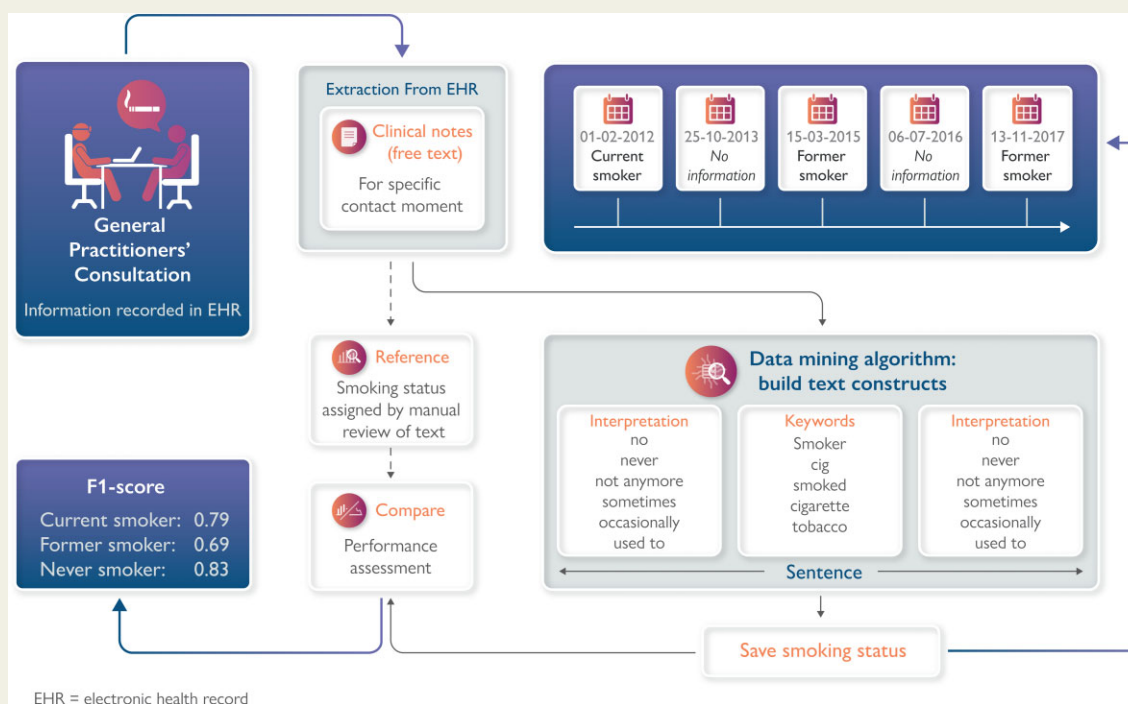
Data mining can successfully retrieve smoking status information from general practice clinical notes with a good performance for classifying ever and never smokers. Differences between general practice and hospital EHRs call for optimization of data mining algorithms when applied beyond a primary development setting.

* Corresponding author. Fax: +31 088 75 68099, Email: a.r.deboer-9@umcutrecht.nl

© The Author(s) 2022. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Graphical Abstract



Keywords

Data mining • Electronic health records • Smoking • Routine clinical data

Introduction

Increasingly, electronic health records (EHRs) are used as a data source for research purposes.^{1–3} Electronic health records, containing routine healthcare data, provide a unique insight into the daily clinical care of all patients, not hindered by any constraints in the form of exclusion criteria or selective participation as could be the case in cohort or trial data.⁴ Information captured in EHRs can be recorded as structured or unstructured data. Clinical notes are an example of unstructured data that contain patient-specific information capturing nuances and clinical reasoning. However, extraction and subsequent interpretation of clinical notes is challenging. There is heterogeneity among clinicians regarding note-taking, including spelling errors, abbreviations, and overall differences in writing style. Yet, by using data mining techniques, it is possible to retrieve information from EHRs including clinical notes.⁵ For example, Brunekreef *et al.*⁶ developed a rule-based text-mining algorithm to identify and characterize patients with systemic lupus erythematosus from clinical notes.

In cardiovascular research and healthcare, having information about smoking status is key, since it is being used to identify those at high risk of cardiovascular disease, or is used as a predictor, a confounder, or a modifier when addressing research questions.⁷ In the Netherlands, most cardiovascular risk management takes place in general practice, and 75% of people visit their general practitioner (GP) at least once a year.⁸ Therefore, general practice EHRs have

the potential to be a valuable source of information, including smoking status information. In general practice EHRs, smoking status can be listed in structured fields using the International Classification of Primary Care (ICPC) code P17 (tobacco dependence). However, P17 is generally reserved for those with extensive tobacco abuse and is lacking in information such as attempts to quit smoking or start smoking again after quitting. Generally, when data are extracted from general practice EHRs for research, smoking information is not provided apart from the P17 code. This despite the fact that more information on smoking behaviour may be found in clinical notes, provides not only more nuance but also longitudinal information, making research into smoking trajectories possible.

First examples of natural language processing algorithms to classify smoking status from clinical records originate from the 'smoking challenge' issued as part of the i2b2 project (Informatics for Integrating Biology to the Bedside),⁹ and since then, more have followed or expanded upon this initiative.^{10–12} Differences between countries and settings (e.g. hospital vs. general practice settings), particularly differences in language, can complicate or make it impossible to reuse or build upon previously developed open-source algorithms. Groenhof *et al.*¹³ developed a rule-based data mining algorithm to retrieve information about smoking status (i.e. never, current, and former smoker) from Dutch hospital EHRs. In this study, we aim to optimize and assess the performance of an existing data mining algorithm in clinical notes in general practice EHRs.

Methods

Data source

We used data from the JGPN. The database contained pseudonymized routine healthcare data extracted from structured and unstructured fields within the EHRs from all patients ($n = 370000$) registered in 72 general practices from the city of Utrecht and its vicinity in the Netherlands.¹⁴ In the Netherlands, all inhabitants (except elderly people dwelling in nursing homes) are obliged to register at a general practice and have access to healthcare, since healthcare insurance is mandatory. General practitioners act as gatekeepers to hospital care and play a key role in cardiovascular risk management. As such, Dutch general practice data are a reliable reflection of the health status of a majority of the Dutch population. The JGPN population is considered representative of the Dutch population with regard to sex and age.¹⁴ However, the data underlying this article cannot be shared publicly due to privacy regulations.

Data collection

We extracted all available unstructured information (clinical notes) from 992 patients randomly selected from the JGPN database, with an oversampling of patients with a cardiovascular history to increase the possibility of finding information about smoking status. The extracted data were randomly divided into a training set ($n = 498$) and a test set ($n = 494$). Medical history is registered within the general practice EHR according to the ICPC codes. We defined the existence of a cardiovascular history if a patient had one of the following ICPC codes registered: K74, K76, K77, K86, K87, K90, K91, and K92. Each time a patient has contact with the GP, this is registered as a contact moment. For each contact moment, clinical notes are registered according to a preset structure, called SOAP. SOAP stands for Subjective, Objective, Assessment, and Plan and is a method of documenting a consultation in the EHR in general practice. For this study, we extracted all SOAP notes from every consultation, date of consultation, ICPC code of the specific consultation, year of birth, sex, and medical history including cardiovascular disease, diabetes (ICPC code T90), and chronic obstructive pulmonary disease (ICPC code R95, R96, R91). For ICPC code definitions, we refer to [Supplementary material online, Table S1](#).

Reference

For both the training and the test sets, each contact moment was assigned a smoking status classification (i.e. 'current smoker', 'former smoker', 'never smoker', or 'no information') using all clinical notes taken during the specific contact moment, which served as a reference for the data mining algorithm. Reference smoking statuses were assigned by manually reviewing (A.R.d.B.) and interpreting clinical notes from 5000 contact moments from 100 patients of the training set. The remaining clinical notes that contained smoking status information were found by searching for (parts of) words that would indicate that smoking information was present, such as 'sm', 'cigarette', 'CVRM' (CardioVascular Risk Management), etc. and were subsequently manually reviewed in its totality and assigned a smoking status. If no smoking status information was detected in the clinical notes, they were assigned the 'no information' classification. The test set was assigned a reference after algorithm optimization had been finalized.

Data mining algorithm

We used the data mining algorithm previously developed by Groenhof *et al.*¹³ and optimized this specifically to be used in general practice EHRs. The data mining algorithm can best be described as a decision

rule model. First, the algorithm mines information on smoking status that is captured in clinical notes for each contact moment separately. Second, for each contact moment, the retrieved information was categorized as either 'current smoker', 'former smoker', 'never smoker', or 'no information'. Free-text fragments from clinical notes were used to build text constructs: first, a keyword word (smoking, smoked, smoker, ...) was selected, then, the surrounding sentence fragments were assessed for interpretation (\pm , sometimes, quit, ...), and the smoking status was finalized ([Figure 1](#)). If a patient was categorized as 'current smoker' or 'former smoker' for a contact moment, that patient could no longer be categorized as 'never smoker' for a contact moment on a later date. The final output of the algorithm will be a list of dates with assigned smoking statuses encompassing a patients' smoking history ([Figure 2](#)).

Only the training set was used to train and optimize the algorithm as developed by Groenhof *et al.* The algorithm was trained using supervised learning, meaning that the manually assigned reference in the training set was available to train the algorithm on. The optimization process consisted of an iterative evaluation of applying the algorithm to the training set, performance assessment, determining which decisions resulted in misclassifications, and adjusting the algorithm until no alterations could be made, which resulted in improvements in classification. After training and optimization, the algorithm was applied in an independent test set to formally assess the performance of the algorithm.

Statistical analyses

The classification performance of the algorithm was assessed by comparing the data mining algorithm assigned smoking statuses with the reference (manually assigned smoking status). First, we assessed the classification performance of the algorithm as developed by Groenhof *et al.* without any optimization in the training set. Second, to show the result of algorithm optimization, we assessed the performance of the optimized algorithm in the training set. Third, we formally assessed the performance of the optimized algorithm in an independent test set. We assessed algorithm performance for each smoking status category separately and the named entity recognition of the keywords associated with smoking by calculating precision (positive predictive value), recall (sensitivity), and F1-score ([Supplementary material online, Box S1](#)). The F1-score expresses both precision and recall in a single measure and is described as the harmonic mean of the algorithm's precision and recall:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

An F1-score of 1 is the best possible result, and a score of 0 is the worst possible result. We performed a performance assessment on a contact moment level and on a patient level. The first level meant the performance of the algorithm to classify a certain contact moment, and the second level meant the performance of the algorithm to classify patients as either 'never smoker' or 'ever smoker' using all information retrieved by the algorithm from all contact moments of that patient. The data mining algorithm was developed in SAS software, Version 9.4 TS1M6. All analyses were performed in R Statistical Software version 3.5.1, Foundation for Statistical Computing, Vienna, Austria.¹⁵

Results

The training set contained clinical notes from 498 patients with a total of 75 712 contact moments, and the median number of contact moments per patient was 108 [interquartile range (IQR) 43–204]. The test set contained clinical notes from 494 patients with a total

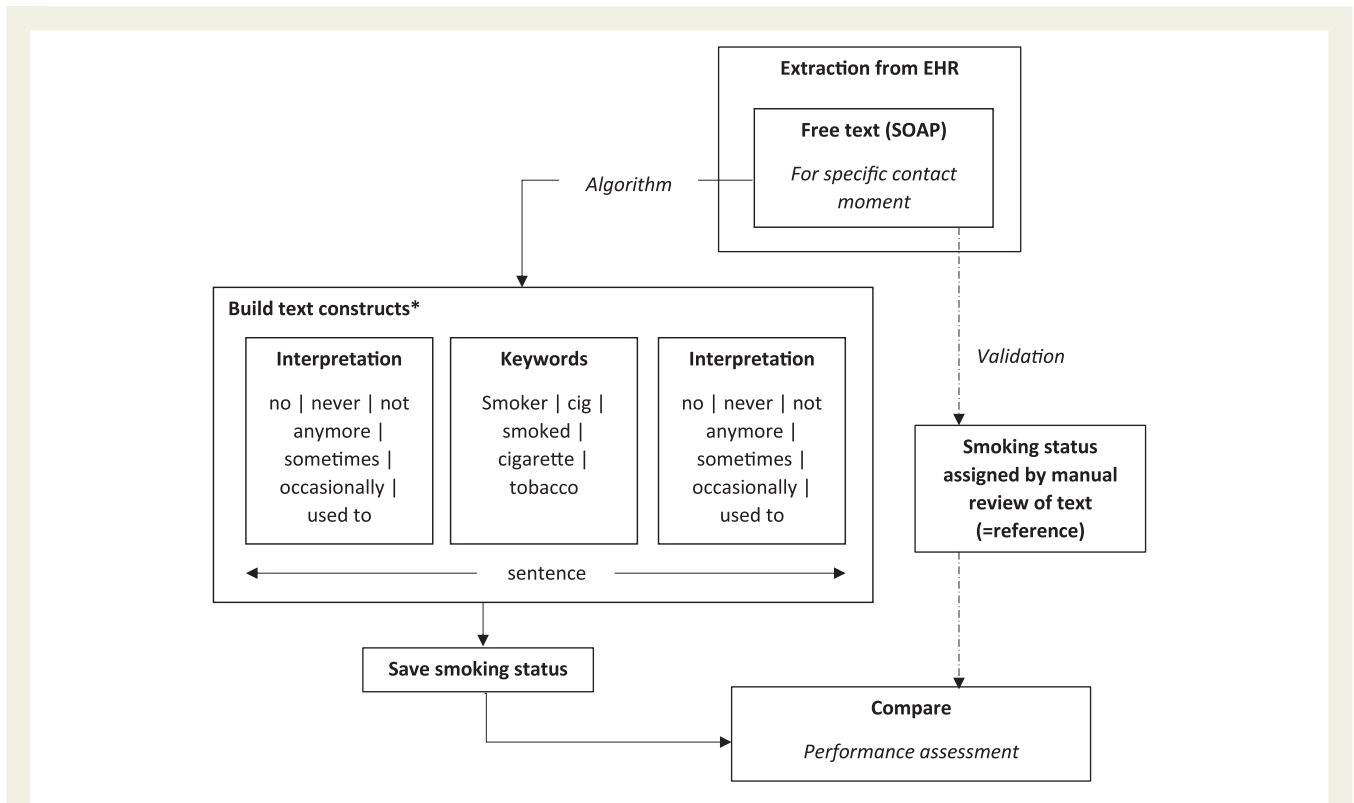


Figure 1 Data mining algorithm and validation process. *Examples of words translated from Dutch. EHR, electronic health record.

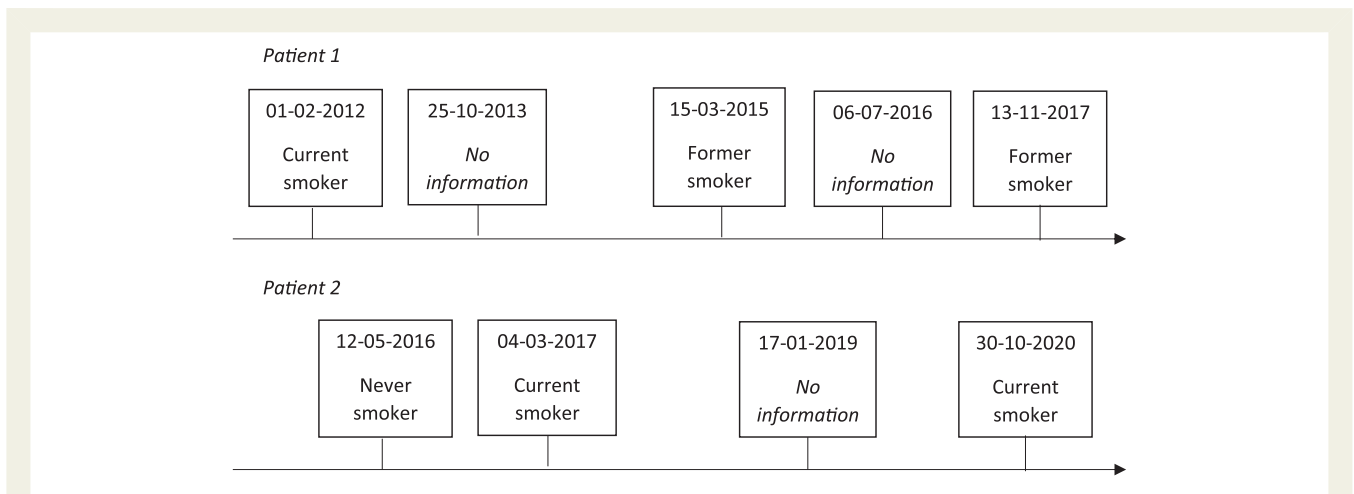


Figure 2 Example data mining algorithm results.

of 78 129 contact moments, and the median number of contact moments per patient was 105 (IQR 47–214). For both the training and the test sets, most of the clinical notes did not contain any smoking status information: in the training set, 1229 (1.6%) of the contact moments of 291 patients (58.4%) contained smoking status information, and for the test set, this was 1457 (1.9%) of the contact moments of 288 patients (58.3%). [Table 1](#) summarizes patients’ characteristics for the training and test sets. Notably, 8.0% of patients in the training set and 10.1% of patients in the test set had a P17 registration.

Algorithm optimization

Before optimization, the algorithm retrieved information about smoking status from 96.3% (1184 of 1229) and performed poorly: F1-score current smoker was 0.72, former smoker 0.43, and never smoker 0.49 ([Supplementary material online, Table S2](#)). After adjustment, the algorithm was able to retrieve information about smoking status from 95.6% (1175 of 1229) contact moments and performance improved considerably: F1-score current smoker was 0.82, former smoker 0.71, and never smoker 0.85 ([Table 2](#)).

Final performance assessment

In the test set, the algorithm was able to retrieve information about smoking status from 89.5% (1304 of 1457) contact moments. [Table 3](#) shows the performance of the test set: F1-score current smoker was 0.79, former smoker 0.69, and never smoker 0.83. If former smokers and never smokers would both be classified as 'current non-smokers', the precision for non-smokers would become 87.4% [95% confidence interval (CI) 84.4–89.8%], recall 73.8% (95% CI 70.4–76.9%), and F1-score 0.80 ([Supplementary material online, Table S3](#)). If current smokers and former smokers would both be classified as 'ever smokers', the precision for smokers would become 93.0% (95% CI 91.2–94.4%), recall 87.4% (95% CI 85.3–89.2%), and

F1-score 0.90 ([Supplementary material online, Table S4](#)). On a patient level, the algorithm was able to retrieve smoking status information from 96.5% (278 of 288) patients who had smoking information in their EHR and correctly identified 98.5% (203 of 206) patients as having no smoking status information in their clinical notes. Note, according to ICPC code P17, only 50 patients would be classified as smokers ([Table 1](#)). The performance to classify a patient as ever smoker was precision 87.9% (95% CI 82.1–92.1%), recall 94.7% (95% CI 89.8–97.4%), and F1-score 0.91 and to classify a patient as a never smoker 98.0% (95% CI 92.2–99.7%), 82.0% (95% CI 73.1–87.8%), and 0.89, respectively ([Table 4](#)).

Table 1 Patient characteristics of the training and test sets

| | Training set (498 patients, 75 712 contact moments) | Test set (494 patients, 78 129 contact moments) |
|--|---|---|
| Contact moments per patient, median (IQR) | 108 (43–204) | 105 (47–214) |
| Men, n (%) | 248 (49.8) | 234 (47.4) |
| Age in years, median (IQR) | 58 (39–74) | 57 (40–73) |
| Follow-up time in years, median (IQR) | 12 (7–23) | 13 (7–24) |
| <i>Medical history</i> | | |
| Cardiovascular history, n (%) | 229 (46.0) | 240 (48.6) |
| Diabetes, n (%) | 97 (19.5) | 102 (20.6) |
| Chronic obstructive pulmonary disease, n (%) | 62 (12.4) | 51 (10.3) |
| Smoking (P17), n (%) | 40 (8.0) | 50 (10.1) |

IQR, interquartile ranges.

Discussion

The current study optimized and applied the data mining algorithm previously developed by Groenhof *et al.*⁹ and assessed its performance in clinical notes from general practice EHRs. We show that portability between these two healthcare settings, however possible, is limited. After optimization, the algorithm was able to retrieve a majority of information on smoking status that was available in clinical notes, and the overall performance in an independent test set was good in classifying patients as ever or never smokers during their lifetime.

Interestingly, during the development stage in which the algorithm was optimized to work in general practice clinical notes, we discovered a distinctive difference between clinical notes in hospital EHRs (in which the algorithm was initially developed) and general practice EHRs. Although some GPs had adopted a systematic writing style in EHRs comparable to the hospital clinical notes (for example: 'alcohol: yes; smoking: yes; BMI: 26'), the majority had adopted a more narrative style of writing (for example: 'wants to quit smoking 6–12 cig, has quit smoking for 9 months, but smokes when feeling bad, partner still smokes, but only outside the house'). This narrative writing style introduces more heterogeneity, especially in words used to give interpretation, and is harder for the algorithm to classify correctly. This resulted in two main algorithm adjustments during the optimization process. First, the hospital EHRs contained a questionnaire that provided information about smoking status in an unambiguous manner that was easy for the algorithm to categorize correctly. The GP EHRs did not contain such questionnaires. Second, the distance in

Table 2 Contingency table training set after algorithm optimization per contact moment

| Smoking status | Reference | | | | Total |
|-----------------------|----------------|---------------|--------------|----------------|--------|
| | Current smoker | Former smoker | Never smoker | No information | |
| Data mining algorithm | | | | | |
| Current smoker | 497 | 74 | 12 | 22 | 605 |
| Former smoker | 61 | 252 | 32 | 0 | 345 |
| Never smoker | 3 | 18 | 202 | 2 | 225 |
| Not classified | 50 | 24 | 4 | 74 459 | 74 537 |
| Total | 611 | 368 | 250 | 74 483 | 75 712 |

Current smoker: precision = 82.1% (95% CI 78.9–85.1%), recall = 81.3% (95% CI 78.0–84.3%), F1-score = 0.82. Former smoker: precision = 73.0% (95% CI 68.0–77.6%), recall = 68.5% (95% CI 63.4–73.1%), F1-score = 0.71. Never smoker: precision = 89.8% (95% CI 84.9–93.3%), recall = 80.8% (95% CI 75.3–85.4%), F1-score = 0.85. Named entity recognition of keywords associated with smoking: precision = 98.0% (95% CI 97.0–98.7%), recall = 93.7% (95% CI 92.1–95.0%), F1-score = 0.96.

Table 3 Contingency table test set per contact moment

| Smoking status | Reference | | | | Total |
|-----------------------|----------------|---------------|--------------|----------------|--------|
| | Current smoker | Former smoker | Never smoker | No information | |
| Data mining algorithm | | | | | |
| Current smoker | 576 | 93 | 30 | 23 | 722 |
| Former smoker | 71 | 265 | 19 | 4 | 359 |
| Never smoker | 2 | 18 | 230 | 0 | 250 |
| Not classified | 87 | 38 | 28 | 76 645 | 76 798 |
| Total | 736 | 414 | 307 | 76 672 | 78 129 |

Current smoker: precision = 79.7% (95% CI 76.6–82.6%), recall = 78.3% (95% CI 75.1–81.2%), F1-score = 0.79. Former smoker: precision = 73.8% (95% CI 68.9–78.2%), recall = 64.0% (95% CI 59.2–68.6%), F1-score = 0.69. Never smoker: precision = 92.0% (95% CI 87.7–94.9%), recall = 74.9% (95% CI 69.6–79.6%), F1-score = 0.83. Named Entity Recognition of keywords associated with smoking: precision = 98.0% (95% CI 97.1–98.7%), recall = 89.5% (95% CI 87.8–91.0%), F1-score = 0.94.

Table 4 Contingency table test set per patient

| Smoking status | Reference | | | Total |
|-----------------------|-------------|--------------|----------------|-------|
| | Ever smoker | Never smoker | No information | |
| Data mining algorithm | | | | |
| Ever smoker | 160 | 19 | 3 | 182 |
| Never smoker | 2 | 97 | 0 | 99 |
| Not classified | 7 | 3 | 203 | 213 |
| Total | 169 | 119 | 206 | 494 |

Ever smoker: precision 87.9% (95% CI 82.1–92.1%), recall 94.7% (95% CI 89.8–97.4%), F1-score 0.91. Never smoker: precision 98.0% (95% CI 92.2–99.7%), recall 82.0% (95% CI 73.1–87.8%), F1-score 0.89.

the number of characters preceding and following the index word used for interpretation needed to be adjusted. Due to the narrative writing style of the GP EHRs, often more smoking-related words were used in one clinical note. The algorithm was adjusted to merge the phrases containing these words and analyse them as one phrase, instead of treating them as separate phrases.

Groenhof et al. evaluated the performance of their algorithm in patients who had both a data mining assigned smoking status and a reference smoking status and found for the current smoker status a PPV (Precision) of 63% (95% CI 58–67%) and a sensitivity (Recall) of 88% (95% CI 83–92%), which translated to a F1-score of 0.73. For non-smokers, this was 98% (95% CI 97–98%), 92% (95% CI 90–94%), and 0.95, respectively.¹³ Compared with that of Groenhof et al., our algorithm performed poorly for non-smokers but better for current smokers. A likely explanation for the poor performance in the non-smoker category is the more narrative style of writing in general practice as compared with hospital EHRs and the lack of information in structured data fields that were available in the EHRs that Groenhof et al. used and that are easier to classify.

The algorithm in our study was able to retrieve information about smoking for almost all patients (a yield of 96.5%) who had smoking status information in their clinical notes. If no information about smoking status was recorded, our algorithm was able to retrieve this information as well (a yield of 98.5%). If both patients with and

without recorded smoking status information would be used for the denominator, the yield of our algorithm (i.e. the percentage of patients from whom a smoking status could be retrieved) would decline to 56.3%. Other studies that focused on the identification of smoking status from EHRs using rule-based models reported a yield between 64 and 94%.^{16–19} The difference in yield can be explained by a difference in study population and rules and regulations between countries. In the Netherlands, smoking status is not recorded by default when a patient is registered in general practice. This is in contrast to studies set for instance in the UK, where registration of smoking status is usual: for example, Marston et al.¹⁷ reported a yield of 84% and Atkinson et al.¹⁸ reported a yield of 94%. Also, two studies reported a kappa (similarity between reference standard and algorithm result) between 0.5 and 0.98.^{17,18} One study trained the algorithm to have a precision of 93% and a recall of 58% but did not assess performance in an independent test set.¹⁶ We expand upon existing rule-based data mining algorithms for smoking status by assessing the final algorithm performance in an independent test set, showing, unsurprisingly, that algorithm performance diminishes when used on new data. Thus, this underlines the importance of assessing algorithm performance on data not used for algorithm development before using such algorithms for scientific purposes or in clinical practice. In this study, we confirm that the portability of data mining algorithms between settings of care is limited, but that existing algorithms can be optimized to perform well in other settings.

Normally, in the extraction of general practice, EHR information on smoking data for research purposes is not included, apart from ICPC code P17. Since P17 is generally reserved for those with extensive tobacco abuse and no structured field is available to indicate if a patient is a non-smoker, this is too limited. In our test set, we could expand the information provided by P17 alone (10.1% of the patients) to a more nuanced information for more patients (56.3%). The implications of the performance of our algorithm (precision and recall) depend on the purpose of the mined data, which could either be scientific or clinical. Supposing the algorithm is used to select eligible patients in a smoking cessation study, the algorithm can be used to make a reliable pre-selection by excluding never smokers without excluding many smokers (high precision for never smoker status) and inviting ever smokers for further selection without missing eligible participants (high recall for ever smoker status). When smoking status information is used to answer aetiological or prognostic

research questions, it can be used as either a determinant, confounder, modifier, or predictor. Again, some information is generally better than no information at all. (i) *Used as determinant*: misclassification may result in an underestimation of the effect of smoking. If current smoker status is used as a determinant, there may remain some misclassifications in favour of former smoker status (precision 79.7%), and some will not be identified at all (recall 78.3%). If ever smoker is used as a determinant, precision (87.9%) and recall (94.7%) will be higher and less misclassification is likely to occur. (ii) *Used as confounder*: especially the current smoker classification is important and misclassification may result in residual confounding. (iii) *Used as effect modifier*: again, differentiation between current smokers and former smokers is important, and the consequence of misclassification may be over- or underestimation of the effect of the exposure in the smoker subgroup. (iv) *Used as predictor*: misclassification may result in an underestimation of the prognostic value. Used in clinical practice, an algorithm can be of great benefit to mine the EHR and use this information in a clinical decision support tool.²⁰ For example, if mined smoking status is used in a cardiovascular risk assessment tool, a misclassification of current smoker status in favour of former smokers may result in a risk overestimation. This is not so worrisome though, since one will detect this during the consultation. If, however, patients are not identified as smokers (recall 78.3% for current smoker status), it could lead to an underestimation of risk and a possible undertreatment. The algorithm can then be used to identify those without any smoking status information and give a notification if a patient visits the general practice so that this information can be acquired during the consultation.

The strengths of this study are the large sample of routine general practice data and performance assessment in an independent test set. Furthermore, we assign smoking status to each individual contact moment and as a result have longitudinal information on smoking per patient. Limitations include the limited generalizability of our algorithm to EHRs in other languages besides Dutch and to domains other than general practice. However, we have shown that an existing algorithm can be optimized to perform in another setting. In a large proportion of patient records (43.7%), no information about smoking status is recorded, which is a reflection of daily clinical practice. Furthermore, the data mining algorithm operates under the assumption that patients could no longer be categorized as 'never smoker' for a contact moment after a previous 'current' or 'former' smoker classification. We assume the first entry to be correct, because we expect that patients are more inclined to trivialize their smoking behaviour. However, from the clinical notes alone, we cannot entirely be sure which entry is wrong. Additionally, the reference standard is formulated by manually scrutinizing and interpreting the clinical notes by one author only. Unfortunately, tests in, for example, urine or blood to confirm smoking status are not available as an objective reference standard. Lastly, the algorithm shows no perfect performance, which is to be expected. The implications of misclassification depend on the purpose of the algorithm, which is further discussed in this section.

Conclusion

Data mining can successfully retrieve smoking status information from general practice clinical notes with a good performance for

classifying ever and never smokers. Differences between general practice and hospital EHRs call for optimization of data mining algorithms when applied beyond a primary development setting.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Acknowledgements

We gratefully acknowledge the contributions of the Julius General Practitioners' Network in providing the data.

Funding

A.R.d.B. and I.V. are supported by the Facts and Figures grant from the Dutch Heart Foundation. S.H. is supported by an Abbott Diagnostics fellowship.

Conflict of interest: None declared.

Data availability

The data underlying this article cannot be shared publicly due to their containing information that could compromise the privacy of included individuals.

References

- Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, van Thiel GJM, Cronin M, Brobert G, Vardas P, Anker SD, Grobbee DE, Denaxas S, Innovative Medicines Initiative 2nd programme, Big Data for Better Outcomes, BigData@Heart Consortium of 20 academic and industry partners including ESC. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *Eur Heart J* 2018;**39**:1481–1495.
- Hemingway H, Feder GS, Fitzpatrick NK, Denaxas S, Shah AD, Timmis AD. Using Nationwide 'Big Data' from Linked Electronic Health Records to Help Improve Outcomes in Cardiovascular Diseases: 33 Studies Using Methods from Epidemiology, Informatics, Economics and Social Science in the ClinicAI Disease Research Using Linked Bespoke Studies and Electronic Health Records (CALIBER) Programme. Southampton, UK: NIHR Journals Library, Programme Grants for Applied Research; 2017.
- Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and pitfalls of electronic health record analysis. *Diabetologia* 2018;**61**:1241–1248.
- Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. *Yearb Med Inform* 2017;**26**:38–52.
- 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**:1007–1015.
- Brunekreef TE, Otten HG, van den Bosch SC, Hoefler IE, van Laar JM, Limper M, Haitjema S. Text mining of electronic health records can accurately identify and characterize patients with systemic lupus erythematosus. *ACR Open Rheumatol* 2021;**3**: 65–71.
- Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, Cooney MT, Corrà U, Cosyns B, Deaton C, Graham I, Hall MS, Hobbs FDR, Lochan ML, Löllgen H, Marques-Vidal P, Perk J, Prescott E, Redon J, Richter DJ, Sattar N, Smulders Y, Tiberi M, van der Worp HB, van Dis I, Verschuren WMM, Binno S, ESC Scientific Document Group. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: the Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J* 2016;**37**: 2315–2381.
- Central Bureau of Statistics. Ongeveer drie kwart bezoekt jaarlijks huisarts en tandarts. <https://www.cbs.nl/nl-nl/nieuws/2013/27/ongeveer-drie-kwart-bezoekt-jaarlijks-huisarts-en-tandarts> (20 December 2021).
- Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008;**15**:14–24.

10. Sohn S, Savova GK. Mayo clinic smoking status classification system: extensions and improvements. *AMIA Annu Symp Proc* 2009;**2009**:619–623.
11. Caccamisi A, Jørgensen L, Dalianis H, Rosenlund M. Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. *Ups J Med Sci* 2020;**125**:316–324.
12. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019;**19**:1–13.
13. Groenhof TKJ, Koers LR, Blasse E, de Groot M, Grobbee DE, Bots ML, Asselbergs FW, Lely AT, Haitjema S, UPOD; UCC-CVRM Study Groups. Data mining information from electronic health records produced high yield and accuracy for current smoking status. *J Clin Epidemiol* 2020;**118**:100–106.
14. Smeets HM, Kortekaas MF, Rutten FH, Bots ML, van der Kraan W, Daggelders G, Smits-Pelster H, Helsper CW, Hoes AW, de Wit NJ. Routine primary care data for scientific research, quality of care programs and educational purposes: the Julius General Practitioners' Network (JGPN). *BMC Health Serv Res* 2018;**18**:735.
15. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018.
16. Wu CY, Chang CK, Robson D, Jackson R, Chen SJ, Hayes RD, Stewart R. 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.
17. Marston L, Carpenter JR, Walters KR, Morris RV, Nazareth I, White IR, Petersen I. 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.
18. Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, Brophy ST. 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.
19. Patel J, Siddiqui Z, Krishnan A, Thyvalikakath T. Leveraging electronic dental record data to classify patients based on their smoking intensity. *Methods Inf Med* 2018;**57**:253–260.
20. Groenhof TKJ, Rittersma ZH, Bots ML, Brandjes M, Jacobs JLL, Grobbee DE, van Solinge WW, Visseren FLJ, Haitjema S, Asselbergs FW. A computerised decision support system for cardiovascular risk management 'live' in the electronic health record environment: development, validation and implementation—the Utrecht Cardiovascular Cohort Initiative. *Neth Heart J* 2019;**27**:435–442.