Combining Structured and Free Textual Data of Diabetic Patients' Smoking Status

Ivelina Nikolova[1(✉)], Svetla Boytcheva[1], Galia Angelova[1], and Zhivko Angelov[2]

[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, 25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
{iva,galia}@lml.bas.bg, svetla.boytcheva@gmail.com
[2] ADISS Ltd., 4 Hristo Botev Blvd., 1463 Sofia, Bulgaria
angelov@adisss-bg.com

Abstract. The main goal of this research is to identify and extract risk factors for Diabetes Mellitus. The data source for our experiments are 8 mln outpatient records from the Bulgarian Diabetes Registry submitted to the Bulgarian Health Insurance Fund by general practitioners and all kinds of professionals during 2014. In this paper we report our work on automatic identification of the patients' smoking status. The experiments are performed on free text sections of a randomly extracted subset of the registry outpatient records. Although no rich semantic resources for Bulgarian exist, we were able to enrich our model with semantic features based on categorical vocabularies. In addition to the automatically labeled records we use the records form the Diabetes register that contain diagnoses related to tobacco usage. Finally, a combined result from structured information (ICD-10 codes) and extracted data about the smoking status is associated with each patient. The reported accuracy of the best model is comparable to the highest results reported at the i2b2 Challenge 2006. These method is ready to be validated on big data after minor improvements.

Keywords: Biomedical language processing · Machine learning · Diabetes risk factors · Preventive healthcare

1 Introduction

Chronic diseases have become epidemiology in last decades and main cause for increasing mortality risks and the rapid increase of health care costs. Recently at the national level was started initiative for development of new technologies and data repositories for retrospective analyses in order to support the health management. In 2015 in Bulgaria a Diabetes Registry (DR) was created automatically [2] with the help of natural language processing techniques applied on the outpatient records (ORs). The ORs are submitted from all kinds of professionals to the Bulgarian National Health Insurance Fund (NHIF), for the period 2012–2014. Diabetes Mellitus is a major cause of cardiovascular diseases,

C. Dichev and G. Agre (Eds.): AIMSA 2016, LNAI 9883, pp. 57–67, 2016.
DOI: 10.1007/978-3-319-44748-3_6

and leading cause of adult blindness, kidney failure, and non traumatic lower-extremity amputations [14], its treatment is costly. Thus the prevention and early diagnostics have crucial importance. Therefore in this project we are focusing on analysis of the risk factors for Diabetes Mellitus and its complications as they are priority tasks of preventive health care.

The ORs in the DR are partially structured, all diagnoses and the data for drugs, only in case they are reimbursed by NHIF are available as XML fields. However the risk factors are mostly encoded in the plain text fields of the document. Here we present results of our work on automatic smoker status identification based on natural language processing (NLP) techniques combined with structured data.

Information Extraction (IE) has proven to be effective technology which provides access to important facts about the patient health and disease development in large volumes of plain text patient records. IE is a matured technology and now widely applied in industrial applications however its application to biomedical data is often in narrow domain only, it is tied to specific languages and medical practices. These particularities hamper the easy transfer of technologies for biomedical text processing between different languages and tasks. Machine learning and rule-based approaches integrated in various hybrid systems are common and with the development of new resources in the field these methods become more and more robust [6]. Most developed are the methods for English medical text processing boosted by the US initiatives for secondary use of medical health records.

The contents of the article is structured as follows: Sect. 2 describes the related studies on the topic, Sect. 3 outlines the materials the study was performed on, in Sect. 4 we brief our methods, the results are presented in Sect. 5 and in Sect. 6 some conclusions are drawn.

2 Related Work

Different aspects of electronic health records analyses have been explored last decade for Bulgarian language. Most comprehensive work was done on hospital discharge letters of patients with endocrinology disorders where some high performance extractors for symptoms, lab test values, diagnoses, and medication [2] are developed. In the recent years these analyses have been extended towards ORs from various practitioners who submit their records to the NHIF. Medications are being extracted and normalized to ATC codes with comparatively high accuracy. Analyses on chronicle diseases comorbidity have also been done [3]. One of the most significant works in this direction is the automatic development of the Diabetes Registry from the outpatient records available in the NHIF, again with the help of natural language processing techniques [2]. Persons with potential health hazards related to family history of Diabetes Mellitus are studied in [12]. The current work is part of a larger project for exploration of the DR, personal history and certain conditions influencing health status.

The most considerable work on automatic smoker status identification from discharge letters was done at the First i2b2 De-identification and Smoking Challenge 2006 [16]. Similarly we limited the scope of our study only to understanding

of the explicitly stated smoking information. The smoker categories were defined for the challenge as follows: (i) past smoker - somebody who quit smoking more than a year ago; (ii) current smoker - somebody who is currently smoking or has quit smoking less than a year ago; (iii) smoker - it is clear that the patient is a smoker but there is no sufficient information to be classified as (i) or (ii); (iv) non-smoker - somebody who never smoked; (v) unknown - no mentions of smoking status in the discharge letter.

Most of the teams which participated in the challenge apply a two step strategy: (*i*) identifying sentences in the records discussing smoker status and (*ii*) classifying only these sentences into the predefined categories. Most often the first step was performed based on trigger terms. The authors report that excluding the irrelevant sentences increased the performance of their algorithms significantly. Similarly in this study we classify only samples which contain trigger terms. The system which achieved highest results on the test set is presented in [4]. They annotated additional data thus increased their training data sample and used linguistic and engine specific features. The latter ones had major contribution to the system performance. They include semantic features such as semantic types of some medical entities - medication, diagnoses, negation and anti-smoking medication. Similarly we introduce in our system semantic features by assigning category to the terms available in our categorical dictionaries. These will be explained in detail in Sect. 4. Aramaki et al. [1] at the second step apply comparison of each sentence with sentences from the training set. The sum of the similarity measures between each extracted sentence and the most similar sentences in the training set is used to determine the smoking status of the extracted sentence. Another systems incorporating rule-based and machine learning approaches also achieved good results [5]. The authors perform an intermediate filtering of records which are not meaningful to the task. Smoker status identification is an important task in automated structuring of patient records and it is still under development for various languages [9].

3 Materials

The DR contains outpatient records in Bulgarian language provided by the Bulgarian NHIF in XML format. The available records for 2014 are nearly 8 mln. for about 462,000 patients. Although the major part of the information necessary for the health management is available as structured fields, some of the important factors for the patient status and the disease development are only available in the free-text sections like anamnesis, status, clinical examination, therapy. All texts are in Bulgarian but contain variety of terms in Latin (in Latin alphabet) or Latin terms transliterated in Cyrillic alphabet. We process raw data that contain many spelling and punctuation errors. Due to the limited number of language resources for Bulgarian and the telegraphic style of the message in the ORs some of the traditional methods for text analysis are not applicable e.g. sentence splitting, dependency parsing etc. Only very focused narrow context information extraction techniques can be helpful in these settings.

Similarly to the i2b2 challenge, in our study we define 4 smoker categories:

- **smoker** - the text explicitly states that the patient has recently smoked (*yes*);
- **past smoker** - the text has evidences about the successful smoking cessation (*ex*);
- **non-smoker** - the text explicitly states that the patient has never smoked (*no*);
- **unknown** - there is no explicit statement in the text regarding the patient's smoking status (*unkn*).

Following the good practices from the i2b2 challenge initially we extract from all ORs only 256 characters concordances around the trigger words: "пуш" (*push*, root of smoke), "цигар" (*cigar*, root of cigarette) and "тютюн" (*tyutyun*, tabacco). This task is performed by BITool [2] over the DR records from 2014. This context is necessary for the human to judge and annotate the data. However when we train our model we strip out only a narrow context of 7 tokens to the left and to the right of the trigger. The OR sections in which these strings occur and are taken in consideration are: anamnesis, patient status, diagnosis, clinical examinations, treatment recommendations. Then we annotated manually some randomly selected 3,092 concordances (Set 1 in Table 1) and additionally add to Set 1 about 200 concordances (Set 2 in Table 1) mainly for more complicated cases of past smokers, that contain rich temporal information about the smoking status progress (Fig. 1). The first example has class "smoker" and the second one - "past smoker". The annotation is performed per record level with the classes explained above. We annotated with current, past or non-smoker only explicit statements about the smoking status. Expressions like "отказва пушенето" (*quits smoking*) we consider *unknown* since they do not state clearly the smoking status at the moment.

Table 1. Class distribution in the annotated data set.

Class	ex	no	yes	unkn
Set 1 concordances	56	2,059	941	37
Set 2 concordances	220	2,066	966	40

In the ORs the smoker status is expressed with various expressions like:

- пушач (*pushach*, smoker) - class **smoker**
- тютюнопушене цигари/ден: 5 (*tyutyunopushene cigari/den: 5*, tabacco smoking cigarettes/day: 5) - class **smoker**
- тютюнопушене цигари/ден: 0 (*tyutyunopushene cigari/den: 0*, tabacco smoking cigarettes/day: 0) - class **non-smoker**
- тютюнопушене(-) (*tyutyunopushene*, tabacco smoking) - class **non-smoker**
- бивш пушач (*bivsh pushach*, past smoker) - class **past smoker**

Fig. 1. Examples for rich temporal information about the smoking status progress.

– пушач до преди 3 мес. (*pushach do predi 3 mes.*, smoker until 3 months ago) - class **past smoker**
– цигарите! (*tsigarite!*, the cigarettes!) - class **unknown**

The distribution of the classes in the annotated data is shown on Table 1. The classes of current smokers (*yes*) and non-smokers (*no*) are considerably bigger than the past smokers (*ex*) and the unknown cases (*unkn*). The imbalance of the data presupposes that the smaller classes will be more difficult to predict. Among them the ex-smokers are of our interest.

4 Method

The workflow of this study is shown on Fig. 2.

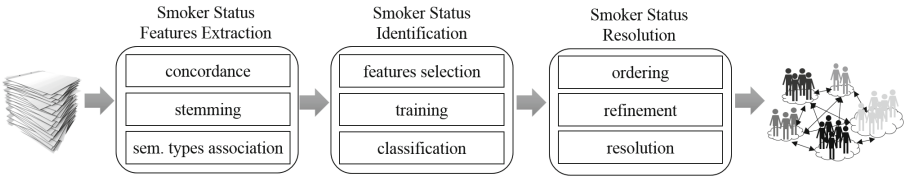

Fig. 2. Workflow.

We perform three stages pipeline. The first stage is responsible for preprocessing of the input data - extracting concordances for the trigger words related to smoking status, stemming these concordances and association of the words with semantic types with the help of 12 vocabularies. In the next stage we perform feature selection and supervised training using manually annotated data. Later this model is refined with additional features extracted from the DR records and the smoker status is being determined.

As explained earlier we focus our work on classifying only ORs containing trigger words signaling smoking. We extract phrases from the free text sections of the OR in the near context of a trigger word. We annotated manually 3,292 of these so called "concordances" and we train a supervised model from the labeled data. The development and training corpus is 66 % of our records and the remaining is test data.

Fig. 3. The vocabulary coverage

We use the following types of features for this task:

- **Linguistic features** - we use the stemmed form [11] of the tokens. Each token stem is an attribute in our feature space except for the stop words. In the latter experiments we also add the verb tense information for the verb smoke.
- **Context features** - these are bigrams, trigrams.
- **Semantic features** - we apply a set of vocabularies which help us to figure out the semantics of the words in the near context. The 12 vocabularies are: *(1)*. Markup terms; *(2)*. Vocabulary of the 100,000 most frequent Bulgarian terms [13]; *(3)*. Generic medical terms in Bulgarian; *(4)*. Anatomical terms in Latin; *(5)*. Generic names of drugs for Diabetes Mellitus Treatment; *(6)*. Laboratory tests; *(7)*. Diseases; *(8)*. Treatment; *(9)*. Symptoms; *(10)*. Abbreviations; *(11)*. Stop words; *(12)*. Negation terms. These are applied in the specified order and the annotations of the latter ones override the previous ones. The categories matched within the concordance are used as features as well as is the number of occurrence of each category. For each concordance is generated single binary vector with bits signaling whether the given attribute is present in the current concordance or not.

The vocabulary coverage is shown on Fig. 3 and Table 2. In the columns are shown the size of each vocabulary (Size), the number of tokens matched in the text by this vocabulary (Tokens), the percentage of tokens in the text matched by this vocabulary (Tokens %), the number of vocabulary entries - types which were matched in the text (Type). The largest coverage has the vocabulary of stop words, then diagnoses, next is the vocabulary of most frequent Bulgarian words followed by the markup words.

Table 2. Lexical profile statistics.

Category	Size	Tokens	Tokens %	Type
1. tags	99	20,684	7.87	29
2. btb	102,730	41,582	15.83	1,051
3. bg med	3,624	1,545	0.59	91
4. term anat	4,382	3,792	1.44	8
5. drugs	154	12	0.01	5
6. lab test	202	18	0.01	5
7. diagnoses	8,444	54,431	20.72	941
8. treatment	339	4,170	1.59	57
9. symptoms	414	4,180	1.59	173
10. abbrev	477	14,404	5.48	83
11. stop words	805	67,153	25.56	166
unknown		50,744	19.32	3,757
TOTAL	121,670	262,715		6,366

The vocabularies lookup and some statistics which helped us for better understanding of the data in means of collocations and terminology are done with AntWordProfiler [10].

5 Results and Discussion

The results shown below are achieved after experiments with various features and instance data size. We narrowed our feature space iteratively starting from a very large space of over 20,000 features. When we restricted the token features only to the ones which appear in 7-token window from the focal term, the attribute space decreased significantly. Then we applied a few rules for filtering out attributes which are not related to smoking and we arrived to about 7,000 attributes in our first experiments. In order to reduce them even more, we applied automatic attribute selection by subset evaluation with default parameters as provided in Weka [7] however the results of the classification in the reduced space were less satisfactory.

Among the algorithms we applied are JRip, LibLINEAR, SMO and SVM with RBF kernel through their Weka implementations or wrappers. In our initial experiments SMO outperformed the other algorithms with 2 to 9 points in F1 for most of the classes therefore the feature engineering phase and final experiments were done with it. SMO is Weka's implementation of John Platt's sequential minimal optimization algorithm for training a support vector classifier. The results reported here are obtained with it only.

We trained our model with 67 % of the data and tested it on the other 33 %. Experiment SMO-1, Table 3 was done on Set1 of the corpus and achieved quite

Table 3. Classification evaluation. SMO-1 - 7,334 attr., SMO-2 - 8,205 attr., SMO-3 - 8,368 attr., SMO-4 - 8,427 attr.

	Precision	Recall	F1	Class
SMO-1	0.92	0.50	0.65	ex
	0.93	0.98	0.96	no
	0.92	0.84	0.88	yes
	0.44	0.36	0.40	unkn
	0.92	0.92	0.92	w. avg
SMO-2	0.89	0.68	0.77	ex
	0.93	0.99	0.96	no
	0.88	0.84	0.86	yes
	0.83	0.33	0.48	unkn
	0.91	0.91	0.91	w. avg
SMO-3	0.85	0.70	0.77	ex
	0.93	0.99	0.96	no
	0.89	0.84	0.86	yes
	0.83	0.33	0.48	unkn
	0.91	0.92	0.91	w. avg
SMO-4	0.88	0.75	0.81	ex
	0.93	0.99	0.96	no
	0.89	0.83	0.86	yes
	0.83	0.33	0.48	unkn
	0.92	0.92	0.92	w. avg

high accuracy for the big classes, however the small classes like *ex* and *unkn* remained hard for guessing. We searched for the reasons not only in the features trained on our development set but also by exploring the data in the DR. Since *ex* is of major importance we analyzed new examples of this class and added them to the corpus. Our expectations were that additional data will lead to improvement of the recognition rate for this class. However the explanations in the ORs of type *ex* are often quite complex and contain a chain of several events related to smoking as shown on Fig. 1. As result the recall indeed improved but the precision has dropped (SMO-2, Table 3). In the next experiments the goal was to improve the precision for *ex* while preserving the achieved accuracy for the big classes. Often past smokers are confused with current ones and less often with non-smokers. Thus some temporality features to distinguish between current and past event have been added. The prepositions which clarify the event smoking were removed from the stop list and added as features to enable bigrams like *smoked until* to enter in the feature set. The results are shown on SMO-3, Table 3. In SMO-4 we added the tense of the verb *smoke* to the feature set. We must mention that the verb *smoke* is used in past tense mostly in records

for *ex-smokers* but also in records for *smoker* such as "was smoking 2 packs a day, now smokes only 10 cigarettes". It appears also in records of *non-smokers* such as "never smoked". Still, introducing this feature lead to higher accuracy for both classes *ex* and *yes*. In these 4 steps we improved the recognition of *ex* with 16 points in F1 while preserving the scores for the majority class *no* and with a minor compromise of 2 points in F1 for class yes.

Table 4. ICD-10 diagnoses for tobacco abuse and NLP. ORs - outpatient records; Ps - patients; non cl. - not classified records; *yes*, *no*, *ex*, *unkn* - manually annotated records with the respective class. Z72.0, F17, Z81.2 - ICD-10 codes.

	ICD-10 only			ICD-10 + NLP			NLP only					
	Z72.0	F17	Z81.2	Z72.0	F17	Z81.2	non cl.	yes	no	ex	unkn	Total
ORs	1,007	23	1	1,113	17	122	820,360	942	2,065	220	39	**825,909**
Ps	609	11	1	968	14	121	457,032	851	1,973	175	36	**461,791**

The instances of class "unknown" are underrepresented in the data set and that is why they are extracted with lower recall. However the precision of the extraction module is comparatively good which means that the features describe well the observed examples. And when dealing with medical data, high precision is a must. Oversampling often helps to increase precision and for real world application it could also be applied. The results we present here are comparable to the ones reported on the i2b2 challenge for smoker status identification from discharge letters in English.

Additional improvement of classification results is possible by taking into account contextualization information. For instance, the concordances extracted from Treatment section refer either to past smoker in case some medication name contains searched key string, or current smoker - in case the searched key string was found in explanations for diet, nutrition and life style recommendation.

In addition to the free text sections of the OR, we analyze also the diagnoses sections. It is not strange that the diagnoses may also contain the triggers we used for extracting the concordances because there are ICD-10 diagnoses [8] like Z71.6 "Tobacco abuse counseling", Z81.2 "Family history of tobacco abuse", P04.2 "Fetus and newborn affected by maternal use of tobacco", T65.2 "Tobacco and nicotine", Z58.7 "Exposure to tobacco smoke", Z72.0 "Tobacco use", Z86.4 "Personal history of psychoactive substance abuse" and F17 "Mental and behavioral disorders due to use of tobacco". Unfortunately these diagnoses are rarely used by professionals, because in the Bulgarian standard for ORs the number of coded diagnoses is at most 5. Another reason for non presence of these diagnoses in ORs is that not all professional encode them explicitly, for instance Ophthalmology. In the DRs for 2014 there are singular ORs that contain codes: T65.2, Z71.6 and Z86.4. Diagnoses P04.2 and Z58.7 are presented in none of the ORs. The majority of markers for smoking status are presented only as free text in the ORs (Table 4). For all patients from the DR we have at least one OR containing

information about their smoking status in 2014. Our ultimate goal is to enrich the patients record in the DR with risk factors information and as of this study - with his/her smoking status. We combine information extracted by NLP techniques and ICD-10 codes (if any). For those patients for who only ICD-10 codes are available - we can resolve the current smoker status as: "smoker" - for T65.2, Z71.6 and Z72.0; "past smoker" - for Z86.4 and "unknown" - for the rest. We can add also status "passive smoker" for Z81.2. And vice versa - for patients without ICD-10 codes for Tobacco use in their ORs we can add the following diagnoses: Z72.0 for "smoker", and Z86.4 for "past smoker". In case both ICD-10 and ORs text contain information about the patient's smoking status - the ICD-10 code can be used for classifier validation. Further investigation of how smoking is influencing health status can be performed on the basis of other diagnosis in the patient's OR and analysis of the temporal information. Similar research was presented in [17], but for ICD-9 codes that include also procedures, however in this study only two classes are considered - ever-smoker and never-smoker.

6 Conclusion

We built a highly accurate model for smoker status identification in Bulgarian outpatient records. Although no rich semantic resources for Bulgarian exist, we were able to enrich our model with semantic features based on categorical vocabularies. The results from this study are comparable to the highest results reported at the i2b2 Challenge 2006. We succeed to improve our model by identifying specific features of the underrepresented classes while preserving the extraction accuracy of the bigger classes. Our next challenge is to apply this model to big data.

There are several risk factors for Diabetes Mellitus that are in the focus of researchers [15] and we plan to continue this work by investigating other potential health hazards like alcohol, drugs, lifestyle and etc. These could be approached with similar means, because ICD-10 provides also diagnoses codes for problems related to lifestyle.

Acknowledgements. This study is partially financed by the grant DFNP-100/04.05.2016 "Automatic analysis of clinical text in Bulgarian for discovery of correlations in the Diabetes Registry" with the Bulgarian Academy of Sciences.

References

1. Aramaki, E., Imai, T., Miyo, K., Ohe, K.: Patient status classification by using rule based sentence extraction and BM25 kNN-based classifier. In: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data (2006)
2. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybern. Inf. Technol. **15**(4), 58–77 (2015)

3. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Mining clinical events to reveal patterns and sequences. In: Margenov, S., Angelova, G., Agre, G. (eds.) Innovative Approaches and Solutions in Advanced Intelligent Systems. Studies in Computational Intelligence, vol. 648, pp. 95–111. Springer, Heidelberg (2016)
4. Clark, C., Good, K., Jezierny, L., Macpherson, M., Wilson, B., Chajewska, U.: Identifying smokers with a medical extraction system. J. Am. Med. Inform. Assoc. **15**, 36–39 (2008)
5. Cohen, A.M.: Five-way smoking status classification using text hot-spot identification and error-correcting output codes. J. Am. Med. Inform. Assoc. **15**, 32–35 (2008)
6. Cohen, K.B., Demner-Fushman, D.: Biomedical Natural Language Processing, vol. 11. John Benjamins Publishing Company, Amsterdam (2014)
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
8. International Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2015/en
9. Jonnagaddala, J., Dai, H.-J., Ray, P., Liaw, S.-T.: A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. In: ACL-IJCNLP 2015, pp. 147–151 (2015)
10. Laurence, A.: AntWordProfiler (Version 1.4.0w) (Computer software). Waseda University, Tokyo, Japan (2014). http://www.laurenceanthony.net/
11. Nakov, P.: BulStem : Design and evaluation of inflectional stemmer for Bulgarian. In: Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics) (2003)
12. Nikolova, I., Tcharaktchiev, D., Boytcheva, S., Angelov, Z., Angelova, G.: Applying language technologies on healthcare patient records for better treatment of Bulgarian diabetic patients. In: Agre, G., Hitzler, P., Krisnadhi, A.A., Kuznetsov, S.O. (eds.) AIMSA 2014. LNCS, vol. 8722, pp. 92–103. Springer, Heidelberg (2014)
13. Osenova, P., Simov, K.: Using the linguistic knowledge in BulTreeBank for the selection of the correct parses. In: Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories, Tartu, Estonia, pp. 163–174 (2010)
14. Rice, D., Kocurek, B., Snead, C.A.: Chronic disease management for diabetes: Baylor Health Care System's coordinated efforts and the opening of the Diabetes Health and Wellness Institute. Proc. (Bayl. Univ. Med. Cent.) **23**, 230–234 (2010)
15. Stubbs, A., Uzuner, Ö.: Annotating risk factors for heart disease in clinical narratives for diabetic patients. J. Biomed. Inform. **58**, S78–S91 (2015)
16. Uzuner, Ö., Goldstein, I., Luo, Y., Kohane, I.: Identifying patient smoking status from medical discharge records. J. Am. Med. Inform. Assoc.: JAMIA **15**(1), 14–24 (2008)
17. Wiley, L.K., Shah, A., Xu, H., Bush, W.S.: ICD-9 tobacco use codes are effective identifiers of smoking status. J. Am. Med. Inform. Assoc. **20**(4), 652–658 (2013)