Medicine®

OPEN

Approaches to text mining for analyzing treatment plan of quit smoking with free-text medical records

A PRISMA-compliant meta-analysis

Hsien-Liang Huang, MD, MSc[a], Shi-Hao Hong, BS[b], Yun-Cheng Tsai, PhD[c,*]

Abstract

Background: Smoking is a complex behavior associated with multiple factors such as personality, environment, genetics, and emotions. Text data are a rich source of information. However, pure text data requires substantial human resources and time to extract and apply the knowledge, resulting in many details not being discovered and used. This study proposes a novel approach that explores a text mining flow to capture the behavior of smokers quitting tobacco from their free-text medical records. More importantly, the paper examines the impact of these changes on smokers. The goal is to help smokers quit smoking. The study population included adult patients that were >20 years old of age who consulted the medical center's smoking cessation outpatient clinic from January to December 2016. A total of 246 patients visited the clinic in the study period. After excluding incomplete medical records or lost follow up, there were 141 patients included in the final analysis. There are 141 valid data points for patients who only treated once and patients with empty medical records. Two independent review authors will make the study selection based on the study eligibility criteria. Our participants are from all the patients that were involved in this study and the staff of Division of Family Medicine, National Taiwan University Hospital. Interventions and study appraisal are not required.

Methods: The paper develops an algorithm for analyzing smoking cessation treatment plans documented in free-text medical records. The approach involves the development of an information extraction flow that uses a combination of data mining techniques, including text mining. It can use not only to help others quit smoking but also for other medical records with similar data elements. The Apriori associations of our algorithm from the text mining revealed several important clinical implications for physicians during smoking cessation. For example, an apparent association between nicotine replacement therapy (NRT) and other medications such as Inderal, Rivotril, Dogmatyl, and Solaxin. Inderal and Rivotril use in patients with anxiety disorders as anxiolytics frequently.

Results: Finally, we find that the rules associating with NRT combination with blood tests may imply that the use of NRT combination therapy in smokers with chronic illness may result in lower abstinence. Further large-scale surveys comparing varenicline or bupropion with NRT combination in smokers with a chronic disease are warranted. The Apriori algorithm suffers from some weaknesses despite being transparent and straightforward. The main limitation is the costly wasting of time to hold a vast number of candidates sets with frequent itemsets, low minimum support, or large itemsets.

Conclusion: In the paper, the most visible areas for the therapeutic application of text mining are the integration and transfer of advances made in basic sciences, as well as a better understanding of the processes involved in smoking cessation. Text mining may also be useful for supporting decision-making processes associated with smoking cessation. Systematic review registration number is not registered.

Editor: Maya Saranathan.

The all authors declare that we have no significant competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

The first and corresponding authors are supported in part by the Ministry of Science and Technology of Taiwan under grant 107-2320-B-002-006.

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

[a] Division of Family Medicine, National Taiwan University Hospital, Zhongzheng Dist., [b] Computer Science and Technology, HeFei University of Technology, Hefei, Anhui Province, [c] School of Big Data Management, Soochow University, Shihlin District, Taipei City, Taiwan (R.O.C.).

[*] Correspondence: Yun-Cheng Tsai, School of Big Data Management, Soochow University, No.70, Linhsi Road, Shihlin District, Taipei City 11102, Taiwan (R.O.C.) (e-mail: pecutsai@gm.scu.edu.tw).

How to cite this article: Huang HL, Hong SH, Tsai YC. Approaches to text mining for analyzing treatment plan of quit smoking with free-text medical records: A PRISMA-compliant meta-analysis. Medicine 2020;99:29(e20999).

Received: 15 January 2020 / Received in final form: 3 April 2020 / Accepted: 29 May 2020

http://dx.doi.org/10.1097/MD.0000000000020999

Abbreviations: ALT = alanine aminotransferase, API = Application Programming Interface, CBC = complete blood count, FTND = Fagerstrom Test for Nicotine Dependence, glucose AC = fasting glucose, GPT = serum glutamic-pyruvic transaminase, HDL = high density lipoprotein, LDL = low density lipoprotein, NAS = Nicotine Addiction Score, NCBI = National Center for Biotechnology Information, NER = named entity recognition, NRT = nicotine replacement therapy, REST = representational state transfer, XML = Extensible Markup Language.

Keywords: association rule mining, decision-making, free-text medical records, quit smoking, text mining

1. Introduction

The disease burden and mortality associated with smoking or secondhand smoke are severe threats to public health. According to the World Health Organization report, tobacco use causes >7 million deaths per year, of which 900,000 are non-smokers exposed to secondhand smoke.(WHO. Influenza. WHO fact sheets No 211. http://www.who.int/mediacentre/factsheets/2003/fs211/en/).[1] Smoking harms almost every organ system in the body and increases the risk of many conditions, including coronary heart disease, stroke, respiratory diseases, gastrointestinal disorders, and cancers.[2] Moreover, the all-cause mortality is about 3 times higher among smokers, and they typically have 10 years lower life expectancy compared with those who had never smoked.[3,4] Therefore, it is essential to help smokers quit for public and personal health.

Smoking is a complex behavior associated with multiple factors such as personality, environment, genetics, and emotions. Some individual characteristics, including decreased control over behavioral impulses and neuroticism, are related to continued smoking, which further results in nicotine dependence.[5–7] Moreover, when smokers attempt to quit smoking, nicotine withdrawal symptoms cause difficulties. Adverse effects, such as agitation, aggression, anxiety, and physical discomfort, are critical factors for smoking relapse.[8,9] Both physical and psychological aspects need to address during cessation periods, and medication use is essential among these smoking cessation interventions. Current practice guidelines recommend the use of medications to ease the transition from smoking to tobacco abstinence in most situations. The most widely used medicines in cessation include nicotine replacement therapy (NRT), varenicline, and bupropion.[10] Nicotine replacement therapy is available as chewing gum, oral and nasal sprays, inhalators, and skin patches. Both varenicline and bupropion are tablets. The effectiveness of these medications differs for specific groups of smokers, and the safety prole or side effects are also different among smokers.[11–14] However, despite the complete and accessible smoking cessation programs aimed to combat this leading cause of preventable illness, previous studies have demonstrated just from 30% to 40% abstinence rates.[10]

Under this backdrop, it is essential to identify factors at the personal level that are associated with difficulties in cessation and to developing tailored interventions that incorporate individual differences.

The free-text data is a rich source of information. However, pure text data requires substantial human resources and time to extract and apply the information, resulting in many details not being discovered and used. Text exploration is a method of sorting out data from the lack of structure and lack of quantitative text data. This method can further combine with various techniques to classify, extract, and streamline data to further meet needs after data exploration. Besides, these techniques can be used to forecast or assist decision making. At present, text-searching techniques are also pervading medical research. One recent example explored forecasting the needs of emergency medical consultations and future hospitalization by text mining and word exploration.[15] Textual exploration has been used to predict and identify cardiovascular diseases,[16] explore factors associated with human papillomavirus (HPV) conditions using textual medical records,[17] assist with drug monitoring by identifying adverse drug event indicators, and in other medical-related applications.[18] The above examples illustrate that textual exploration has value and potential in different disciplines within medical research. Nevertheless, there are few studies in the field of smoking cessation that use text exploration methods.

Therefore, the paper explains an algorithm for analyzing smoking cessation treatments documented in free-text medical records. The data were real-world patients recruited from smoking cessation clinics in a medical center using different medications for smoking cessation. The text mining flow for free-text medical records captures the behaviors of smokers and explores the impact of these changes on smokers trying to quit. Additionally, the paper is a shift in emphasis on "big data" to "small data" analytics as healthcare systems focus on leveraging existing data to improve clinical and operational processes. Our algorithm works for small data, design according to the data analysis process, and easily build a data decision analysis system. The approach involves the development of an information extraction flow that uses data mining techniques that can be used not only to quit smoking but also for other medical records with similar data elements.

The study aimed to use text mining approaches to analyze free-text medical records in smoking cessation clinics. The result might help build person-centered approaches enhancing smoking cessation and inspire the use of text mining in other medical fields.

2. Methods

The study performs based on the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guidelines. The study is a meta-analysis, an analysis with secondary processing. Thus, ethical approval was not necessary for the study.

The summary form was recorded and provided by health instructors from the Family Medicine Department of National Taiwan University Hospital. It contains fields for assessing the smoking status of a patient. The medical records obtained from the outpatient system of the National Taiwan University Hospital. They contain all the medical information before the participation in the smoking cessation clinic. They are including the main symptoms, the results of the examination, and the diagnosis results. They also contain the names of drugs prescribed and tests used for diagnosis, among other information.

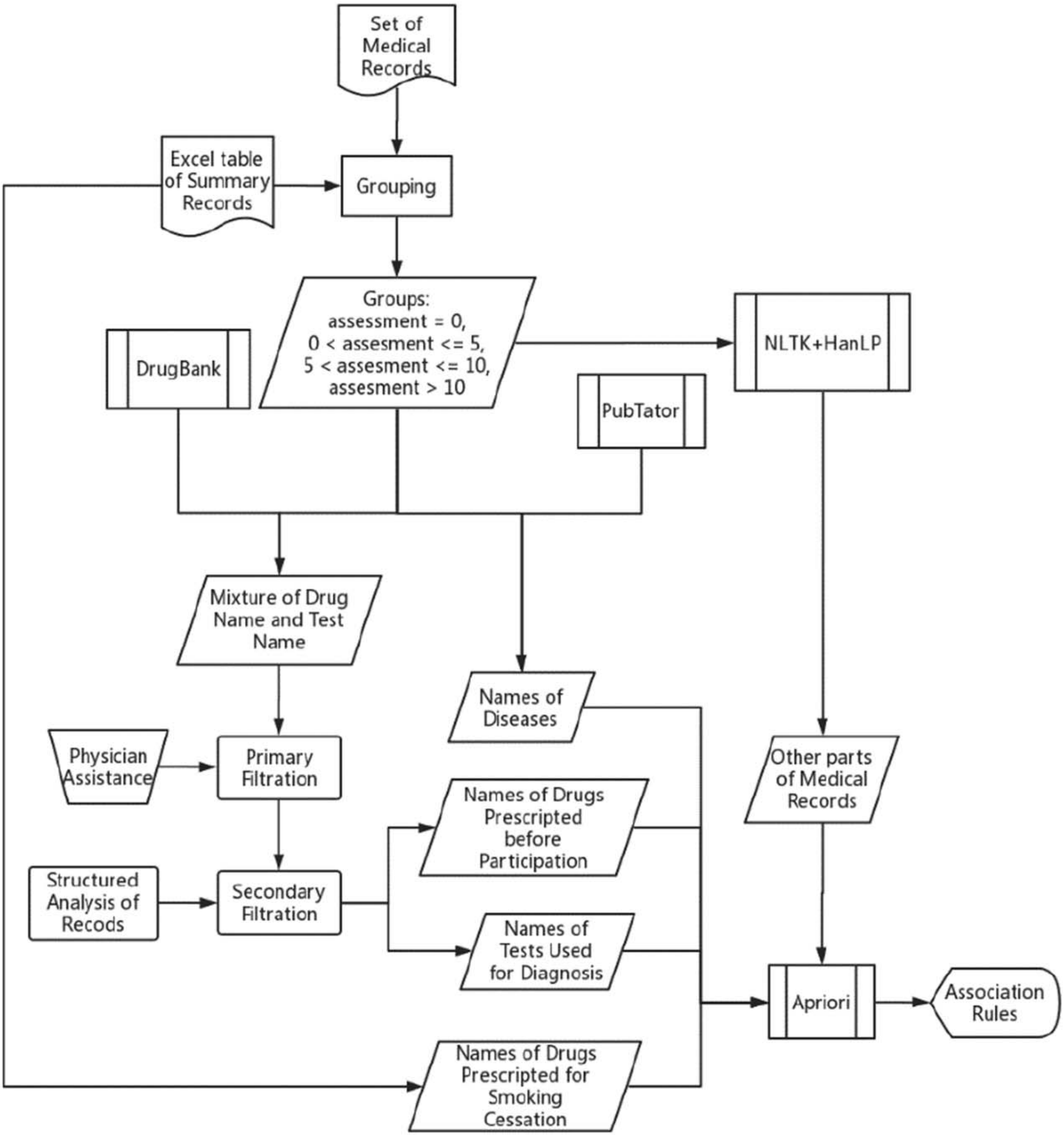
Set of Medical Records
Excel table of Summary Records
Grouping
Groups: assessment = 0, 0 < assesment <= 5, 5 < assesment <= 10, assesment > 10
DrugBank
NLTK+HanLP
PubTator
Mixture of Drug Name and Test Name
Names of Diseases
Physician Assistance
Primary Filtration
Other parts of Medical Records
Names of Drugs Prescripted before Participation
Structured Analysis of Recods
Secondary Filtration
Names of Tests Used for Diagnosis
Apriori
Association Rules
Names of Drugs Prescripted for Smoking Cessation

Figure 1. Text mining flow to discover relationships based on medical records.

2.1. Development of the text mining algorithms

The text mining process with named-entity recognition (NER) and word frequency counting was applied to explore the relationships among texts and smoking cessation outcomes from medical records. Figure 1 shows the overall framework of the analysis as follows.

1. Patient grouping and outcome ascertainment.
2. Annotation with named-entity recognition (NER) and word frequency.
3. Word frequency exploration, analysis, and finding associations.

First, with the doctor's assistance, we identified and extracted useful words from the medical records of different patient smoking cessation groups, such as prescription drugs, test indicators, and disease names. These words accurately describe various aspects of the history of different patient groups. Then, data exploration methods were used to observe correlations between these medical records. Thus, when a new patient joins, we could compare their medical records with the correlation graph and determine which modes they resemble. Finally, the physician would consider this in the diagnosis to determine a definitive medical treatment.

Traditional smoking cessation analysis has only the 2 extreme discrete concepts of success and failure, but we believe that it is quantifiable to a certain extent. A patient's progress in smoking addiction before and after the start of 4 courses can reflect the success of smoking cessation. In particular, the National Taiwan University Hospital records both the initial addiction index and the status of smoking in each course of treatment, which provided us with a basis for quantitative analysis. In this module, the collection of medical records divides into 4 clusters based on assessment scores: success, close to victory, regular, and failure.

2.2. Ascertainment of outcomes

The level of nicotine dependence of patients in the first encounter to cessation clinic reflects the difficulties of smoking cessation, and the smoking status at the last visit of the clinic indicated the success rate of abstinence. Our primary outcome definesv as:

$$\text{assessment} = \frac{\text{LastTimeNum}}{\ln \ln (e + \text{FTND})}$$

where FTND is the Fagerstrom Test for Nicotine Dependence scores recorded in the first visit to the smoking cessation clinic, and LastTimeNum is the average number of cigarettes smoked in the last course of treatment. The outcome is positively correlated to the average daily smoking habits of the last survey and inversely proportional to the nicotine addiction score of the first survey. This formula reflects the following 3 aspects:

1. The lower the assessment score, the better the smoking cessation effect.
2. If the average daily smoking habit of 2 patients are the same, the patient with high FTND will work better.
3. If LastTimeNum=0, the smoking cessation is successful and the assessment is 0.

The value of FTND is an integer from 1 to 10. Since the minimum value of FTND is 1, ln(e + NAS) must be >1.

2.3. Data source and study samples

The study population included adult patients that were >20 years old of age who consulted the medical center's smoking cessation outpatient clinic from January to December 2016. A total of 246 patients visited the clinic in the study period. After excluding incomplete medical record or lost follow up, there were 141 patients included in the final analysis. There are 141 valid data points for patients who only treated once and patients with empty medical records. The summary table records information about smoking cessation, including the nicotine addiction index at the time of the patient's first visit, the calculated Nicotine Addiction Score (NAS), in scoring the Fagerstrom Test for Nicotine Dependence, the prescription medication for each course of treatment, the amount of smoking, and some regular physical condition indicators. The medical records contain 4 parts of free text, including subjective symptoms of patients, objective evaluation by physicians, diagnosis of the patients with current smoking status, and further management plans for the patients. The objective evaluation of the patients includes the Fagerstrom Test for Nicotine Dependence (FTND), the prescribed medication during each clinic visit, the amount of smoking, laboratory results for the comorbidity diseases of the patient, and side effects of the medication.

The study procedure approved by the National Taiwan University Hospital Research Ethics Committee, and all procedures are following the Helsinki Declaration. The ethical committee waived informed consent due to the retrospective design and the complete anonymity of the data.

2.4. Natural language processing with NER

The NER identifies entities with specific meaning in the texts. The Stanford NER targets person, organization, location, etcetera.[19,20] We thus use the NER to extract the patient's features from the free-text medical records after grouping patients according to outcomes. The free-text medical records contained subjective symptoms, objective evaluation including many laboratory tests, diagnosis of the patients with both the standard International Statistical Classification of Diseases and Related Health Problems, and commonly used terms by physicians in the field, and management plans about further medication and medical tests. Under this background, we use physician annotation combined with other tools for the NER.

The first step in the data exploration was to analyze the structure of the data and discuss it with experts in the professional field to successfully extract useful information from the messy, raw data. For this strategy, the original medical record was the raw data, and the analyzable information was the diseases diagnosed, the drugs prescribed, and the tests required. Extracting different types of information requires various tools.

1. The DrugBank (https://www.drugbank.ca/) is a bioinformatics and chemical informatics database provided by the University of Alberta and has Extensible Markup Language (XML) files available for download. The latest version (version 5.1.2) includes detailed information on 11,924 drugs, which can be used to help identify medication in the texts: drug names and test names. We took the sentences in Records 4, compared the word set with DrugBank, and displayed the resulting word length bar graph.
2. The Pubtator is NCBI's web tool[21] that can annotate medical papers in Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/). There are 5 categories: Gene, Disease, Chemical, Species, and Mutation. It also provides a RESTful API, allowing us to process our raw medical records.
3. The NLTK (https://www.nltk.org/) is used for context analysis to identify negative type sentences and inaccurate phrase descriptions in medical records, such as allergies to a drug or the family having a disease. Thus, it helps to eliminate these noise disturbances.
4. The HanLP (https://github.com/hankcs/HanLP) is used to identify and analyze Chinese doctors' orders.

Two physicians who specialize in the smoking cessation field also helped the process of natural language processing. They help disambiguate some terminologies such as medication with a different brand name, generic name, and physicians commonly used terms in free texts records.

2.5. Analysis and finding associations

One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. The Apriori is a seminal algorithm for finding frequent itemsets using candidate generation.[22] It characterizes as a level-wise complete search algorithm using anti-monotonicity of item sets,

Table 1

The example transactions from the DrugBank NER for association rules analysis.

Transactions (patients)	Items (smoking cessation medication)
1	HarnalidgeD_MEDICINE; BerotecMA_MEDICINE; Spiriva Respimat SolnForInh_MEDICINE; Relvar Ellipta Inh; Powder_MEDICINE; Medicon_MEDICINE
2	(DC) Smile Orabase_MEDICINE; Striverdi Respimat Soln for Inh_MEDICINE; Musco_MEDICINE; HarnalidgeD_MEDICINE; PARAMOL_MEDICINE
3	75 Plavix_MEDICINE; BokeyEMcap_MEDICINE; Chewing Gum_SMOKE
4	Augmentin_MEDICINE; Actein Effervescent Tablet_MEDICINE; Chewing Gum_SMOKE
5	PARAMOL_MEDICINE; Naposin_MEDICINE; Lexotan_MEDICINE; C.B. Strong Oint_MEDICINE; Zalain Cream_MEDICINE
6	Salazine EC_MEDICINE; Ultracet_MEDICINE; Inderal_MEDICINE; Nootropil_MEDICINE; Fluitran_MEDICINE
7	1.25 Concor_MEDICINE; COZAAR_MEDICINE; Nitrostat Sublingual_MEDICINE; Brilinta_MEDICINE; BokeyEMcap_MEDICINE

that is, if an itemset is not frequent, none of its supersets are ever frequent.[23]

We make a transaction for each patient's historical medication record and smoking cessation medication. Table 1 shows the information extracted from Records_4. Each row is a patient. Transactions are used as input data to the Apriori algorithm to get frequent itemsets association rules. Association rules analysis is a technique to uncover how items associate with each other. An association rule has 2 parts, antecedent and consequent, and each one has at least 1 feature. The antecedent is also called the left-hand side, and the consequent is called the right-hand side. If there is a rule X to Y, it can denote as $X \rightarrow Y$.

The threshold of the parameters in the Apriori set and the number of rules produced show in Table 2. The more significant the support value, the more likely it is to form a standard set. The paper sorts the above association rules according to the support value, and then draw the relationship diagrams of the first 30 rules. Physicians performed the final interpretation of the associations drawn revealed from the Apriori associations analysis for any clinical application.

After grouping, it is necessary to perform NER analysis of the medical records. Note that NER identifies entities with specific meaning in the text. Stanford NER targets person, organization, location, etc. We thus use NER to extract the patient's features from the free-text medical records. Some common entities include drugs that have been taken by a patient and the chemical tests performed that reflect the physiological characteristics of the patient before they participated in the smoking cessation clinic. It is reasonable to use these as keywords for text analysis. This module uses a variety of tools and methods for the NER.

The DrugBank is a bioinformatics and chemical informatics database provided by the University of Alberta and has XML files available for download. The latest version (version 5.1.2) includes detailed information on 11,924 drugs, which can be used to help identify drug names and test names. We took the sentences in Records_4, compared the word set with the DrugBank, and displayed the resulting word length bar graph. According to the words in the medical records, the doctor judged whether the name referred to drugs or tests, and we summarized the fixed structure of the 2 nouns in the medical records. For example, consider the case where the drug name was the starting position of a specific horizontal line, and the numbers terminated or Percentage, and the term "Health Insurance" appeared on this line. We would rescan the medical records, this time without looking at the sentence, but looking for the structure. After recognizing the test name, we added SMOKE, drug name, and MEDICINE for follow-up discussion in the non-medical field.

Table 2

The threshold and the parameters from final named-entity recognition for the Apriori association rules analysis.

Group	Support	Confidence	Min-length	Max-length	Rule number
Med_class1	0.05	0.8	2	50	5132
Med_class2	0.05	0.8	2	50	299
Med_class3	0.05	0.8	2	50	1035
Med_class4	0.1	0.8	2	50	38
Check_class1	0.2	0.8	2	50	706
Check_class2	0.2	0.8	2	50	10,732
Check_class3	0.2	0.8	2	50	286
Check_class4	0.2	0.8	2	50	3319

The vertical axis of the word frequency map is the number of times the recognized word from the horizontal axis appears in the set, and the single medical record appeared multiple times without repeated calculation. As can be seen from Fig. 2, the DrugBank can identify the name of the drug such as Amoxicillin and mix with the test name such as Corn.

One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. The Apriori is a seminal algorithm for finding frequent itemsets using candidate generation.[22] It is characterized by a level-wise complete search algorithm using anti-monotonicity of item sets, that is, if an itemset is not frequent, none of its supersets are ever frequent.

We made a transaction for each patient's historical medication record and smoking cessation medication. Table 1 shows the information extracted from Records_4. Each horizontal row is a patient. Association rules analysis is a technique to uncover how items are associated to each other. This form of information is acceptable for the Apriori. If there is a pair of items, X and Y, that are frequently taken medication together.

3. Results

A total of 246 patients visited the clinic in the study period. After excluding incomplete medical records or lost follow up, there were 141 patients included in the final analysis. Figure 3 shows the distribution of the 141 patients across assessment categories, defined as follows:

1. There are 74 success cases in Recours_1 (assessment=0).
2. There are 35 patients closed to victory cases in Recours_2 (0<=assessment<5).
3. There are 21 normal cases in Recours_3 (5<=assessment< 0).
4. There are 11 failure cases in Recours_4 (assessment>=10).

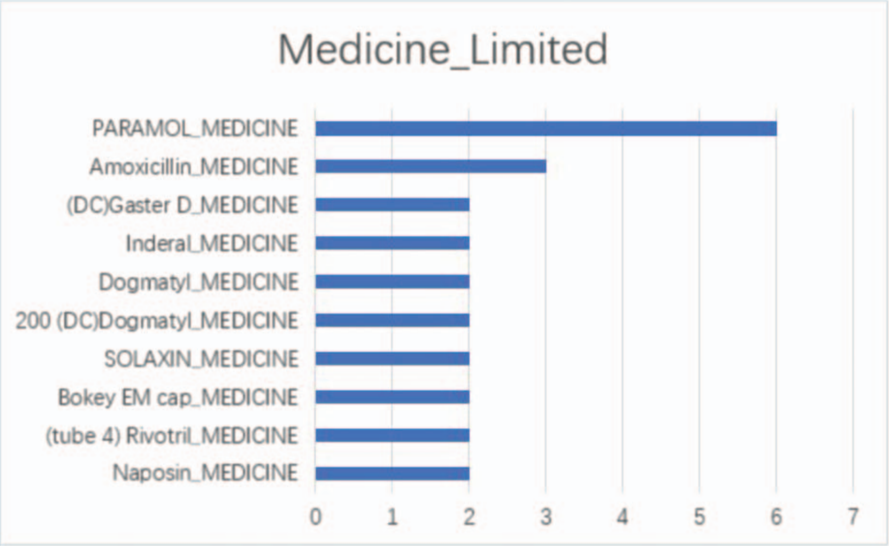

Figure 2. The process of natural language processing with named-entity recognition (NER).

The process of natural language processing with named-entity recognition (NER) shows in Figs. 2 and 4. For example, these 2 histograms were analyzed and presented by the patients in the failed group. Figure 3 is the word frequency counting histogram of the medication in these patients' medical free texts using only the DrugBank database for NER. After using other tools such as the PubMed, the Pubtator, and physicians' annotations, the final word frequency counts histogram demonstrate in Fig. 4.

After creating word frequency counts from the NER, the Apriori algorithm uses for finding the association with the smoking cessation outcomes. When using the Apriori algorithm, the threshold the parameters of the Apriori set and the number of rules produced shows in Table 2. Figure 5 demonstrates the association of medications among the patients who failed to achieve sustained cessation. There was an apparent association between nicotine replacement therapy (NRT) and medications mostly used in patients with psychiatric illnesses such as Inderal, Rivotril, Dogmatyl, and Solaxin. The strong associations among the smokers who succeed in smoking cessation between medication and laboratory items present are in Fig. 6. The items in the association figure were mostly blood and urine laboratory tests ordered by physicians. The most apparent associations of triglyceride, glucose (AC), cholesterol, ALT (GPT), LDL, and HDL reveal in these patients.

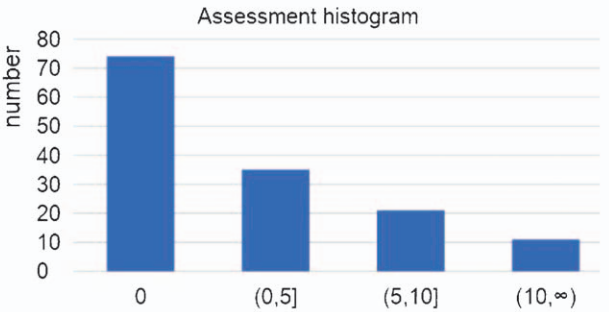

Figure 3. The distribution of patients across assessments categories.

No smoking cessation medication was identified to have healthy relationships in the medical records. Figure 7 presents the Apriori associations of smoking cessation medication with laboratory items in the patients failing in smoking cessation. The laboratory tests, including HDL, LDL, and HbA1c, link to the smoking cessation medications of combination NRT with Nicotinell TTS and Chewing Gum. Other associations of CBC, glucose (AC), cholesterol, triglyceride, ALT (GPT), and creatinine also reveal in the figure.

4. Discussion

The results demonstrated an algorithm for analyzing free-text medical records in the smoking cessation clinic. The data were real-world patients recruited from smoking cessation clinics in a medical center using different medications for smoking cessation. The text mining ow for free-text medical records captures the behaviors of smokers and explores the impact of these changes on smokers trying to quit. Additionally, the paper is a shift in emphasis on "big data" to "small data" analytics as healthcare systems focus on leveraging existing data to improve clinical and operational processes. The approach involves the development of an information extraction flow that uses data mining techniques that can be used not only to quit smoking but also for other medical records with similar data elements. The result might help build person-centered approaches enhancing smoking cessation and inspire the use of text mining in other medical fields.

The Apriori associations from the text mining revealed several important clinical implications for physicians during smoking cessation. For example, Fig. 5 shows an apparent association between nicotine replacement therapy (NRT)

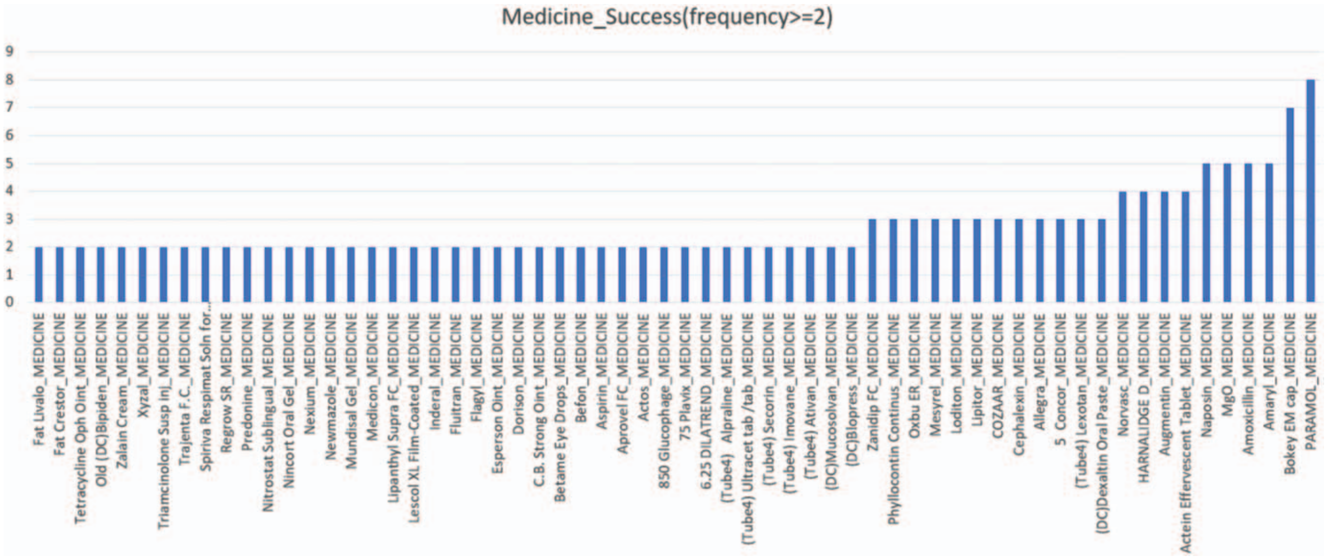

Figure 4. The final word frequency counts histogram.

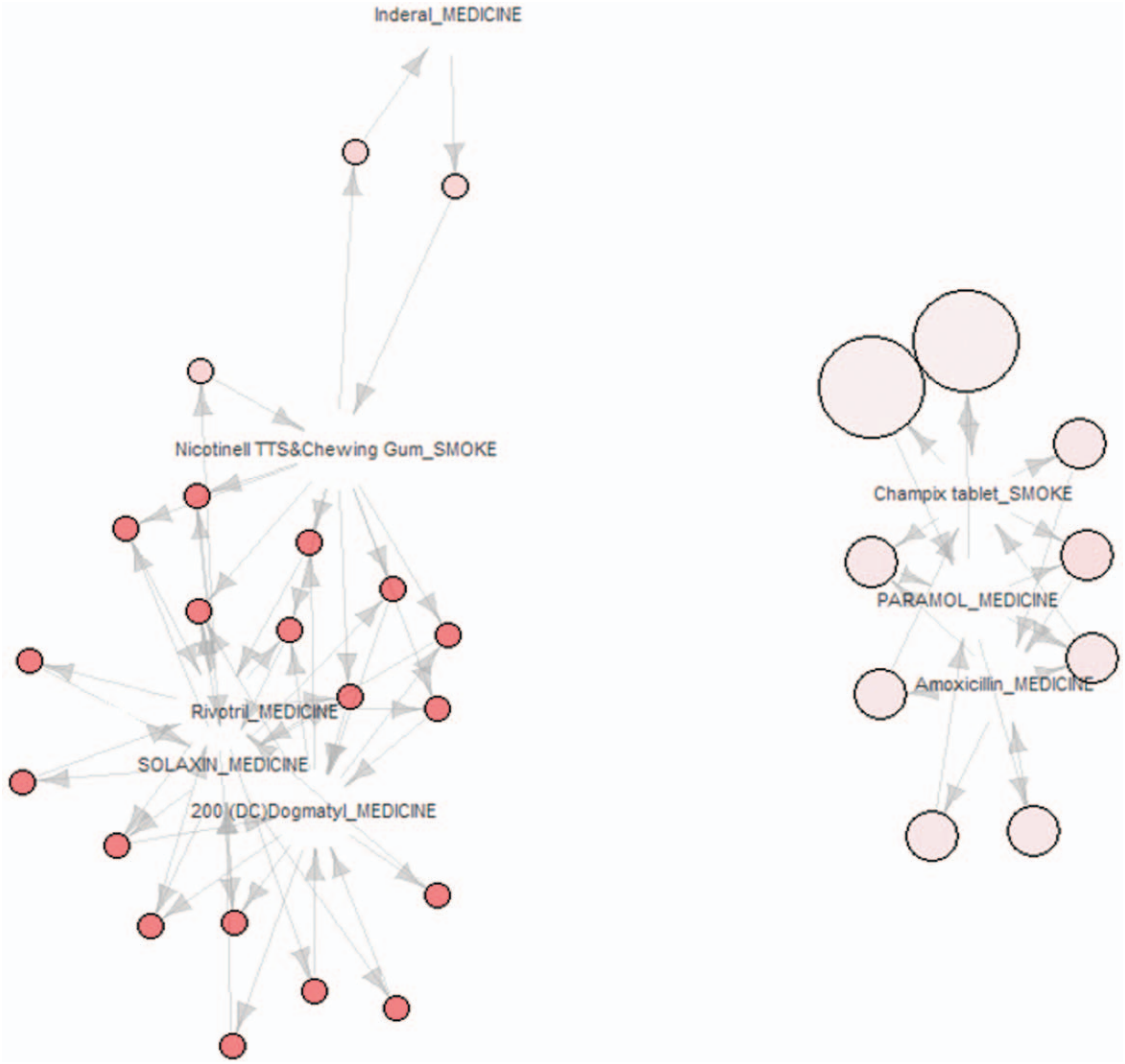

Figure 5. The association of medications among the patients who failed to achieve sustained cessation.

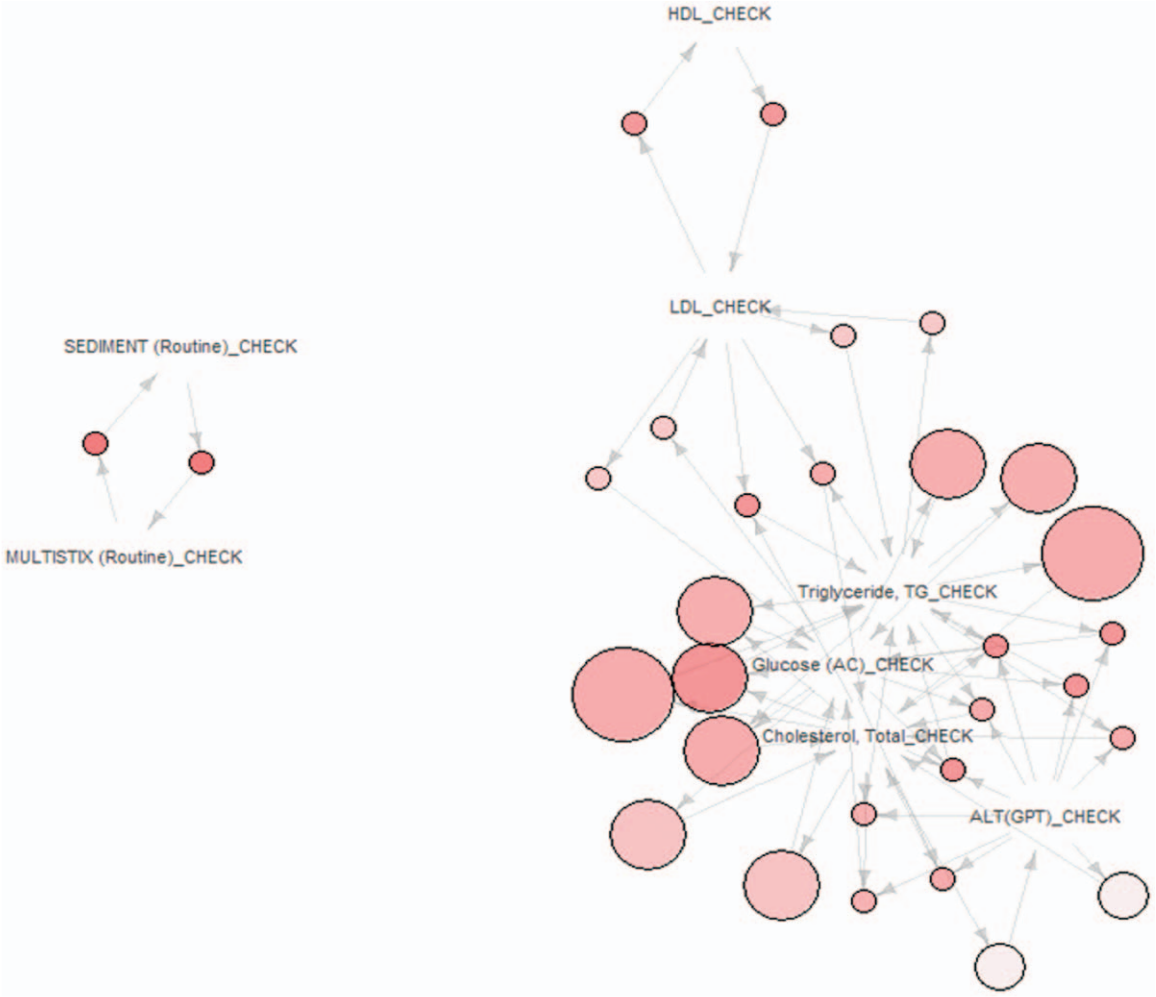

Figure 6. An apparent association between nicotine replacement therapy (NRT) and other medications such as Inderal, Rivotril, Dogmatyl, and Solaxin. Inderal and Rivotril use in patients with anxiety disorders as anxiolytics frequently.

and other medications such as Inderal, Rivotril, Dogmatyl, and Solaxin. Inderal and Rivotril frequently use in patients with anxiety disorders as anxiolytics. Dogmatyl is an antipsychotic that may be used to relieve the emotional symptoms of patients with anxiety. Solaxin is a muscle relaxant also applied to ease the stiffness of muscles in anxiety patients. The difficulties of quitting for smokers with psychiatric disorders discussed in previous studies. Moreover, we concern about the neuropsychiatric safety of the varenicline and bupropion in smoking cessation. That might explain why these 2 medications are not present in the association figure.

Laboratory exams are essential for effective control in chronic illnesses since the results of these blood tests help physicians appropriately adjust medications. Patients with good adherence to medical orders follow physicians' instructions to receive blood tests regularly and thus achieve better outcomes. We may assume that the presence of laboratory exams in patients with successful smoking cessation means that they received blood tests frequently, indicating good adherence to medical advice. The associations in Fig. 6 may implicate that smokers who followed physicians' suggestions during cessation periods were more successful in overcoming the difficulties associated with quitting smoking.

Figure 7 shows that the laboratory tests, including HDL, LDL, and HbA1c, link to the smoking cessation medications of combination NRT with Nicotinell TTS and Chewing Gum. High levels of HDL, LDL, and HbA1C may indicate that patients had hyperlipidemia or diabetes mellitus and needed regular follow-up with blood tests. The rules associating with NRT combination with blood tests may imply that the use of NRT combination therapy in smokers with chronic illness may result in lower abstinence. Further large-scale surveys comparing varenicline or bupropion with NRT combination in smokers with a chronic illness is warranted.

The Apriori algorithm suffers from some weaknesses despite being transparent and straightforward. The main limitation is the costly wasting of time to hold a vast number of candidates sets with frequent itemsets, low minimum support, or large itemsets.

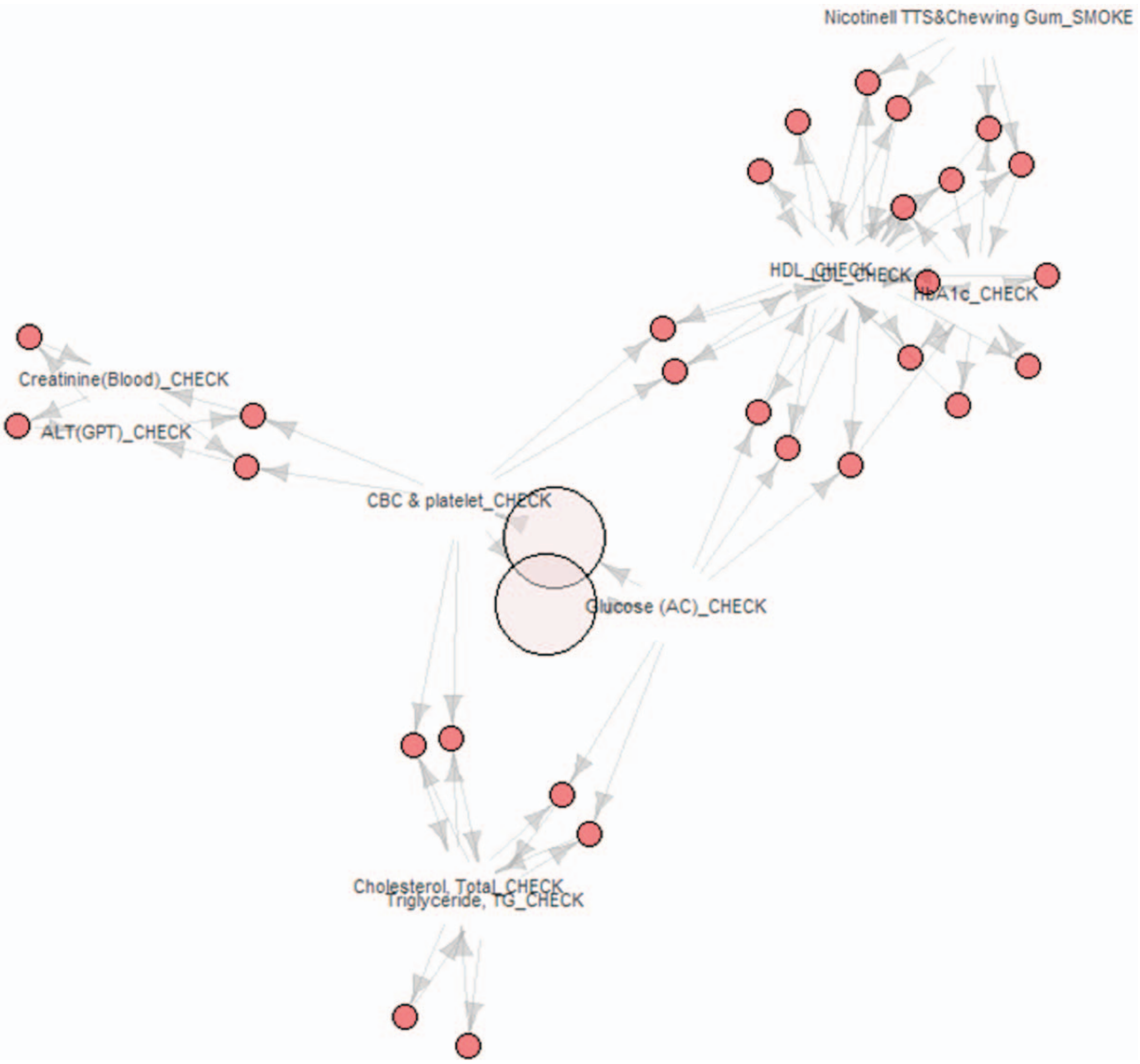

Figure 7. The Apriori associations of smoking cessation medication with laboratory items in the patients failing in smoking cessation.

5. Conclusions

The importance of person-centered strategies for smoking cessation is well supported from our pilot text mining study using free-text medical records. Smokers with psychiatric disorders or chronic illness may need different medications to help quit smoking, which agrees with previous clinical trials. The results here demonstrate that text mining could help clinicians explore important factors in influencing smoking cessation.

A large amount of scientific and medical information is now available. However, there is an inherent problem with such unstructured collection of data: selective recovery and interpretation are practically impossible for a professional using classical methods. In this setting, text mining acquires fundamental relevance.

Text mining and data science already play a significant role in other areas of precision medicine. In our paper, the most obvious areas for the medical application of text mining are the integration and transfer of advances made in basic sciences, and a better understanding of processes to aid in smoking cessation. Text mining may also be useful in the decision-making processes for supporting smoking cessation.

Acknowledgment

The authors would like to thank all the patients that were involved in this study and the staff of Division of Family Medicine, National Taiwan University Hospital.

Author contributions

Data curation: Hsien-Liang Huang.
Investigation: Hsien-Liang Huang.
Methodology: Yun-Cheng Tsai and Shi-Hao Hong.
Supervision: Yun-Cheng Tsai.
Writing – original draft: Yun-Cheng Tsai and Shi-Hao Hong.

References

[1] Öberg M, Jaakkola MS, Woodward A, et al. Worldwide burden of disease from exposure to second-hand smoke: a retrospective analysis of data from 192 countries. Lancet 2011;377:139–46.

[2] Garrett BE, Dube SR, Trosclair A, et al. Centers for Disease Control and Prevention (CDC)Cigarette smoking—United States, 1965–2008. MMWR Surveill Summ 2011;60:109–13.

[3] Thun MJ, Carter BD, Feskanich D, et al. 50-year trends in smoking-related mortality in the united states. N Engl J Med 2013;368:351–64.

[4] Jha P, Ramasundarahettige C, Landsman V, et al. 21st-century hazards of smoking and benefits of cessation in the united states. N Engl J Med 2013;368:341–50.

[5] Carim-Todd L, Mitchell SH, Oken BS. Impulsivity and stress response in nondependent smokers (tobacco chippers) in comparison to heavy smokers and nonsmokers. Nicotine Tob Res 2015;18:547–56.

[6] Buczkowski K, Basinska MA, Ratajska A, et al. Smoking status and the five-factor model of personality: results of a cross-sectional study conducted in Poland. Int J Environ Res 2017;14:126.

[7] Bares CB, Dick DM, Kendler KS. Nicotine dependence, internalizing symptoms, mood variability and daily tobacco use among young adult smokers. Addict Behav 2018;83:87–94.

[8] Bold KW, Witkiewitz K, McCarthy DE. Multilevel factor analysis of smokers' real-time negative affect ratings while quitting. Psychol Assess 2016;28:1033–42.

[9] Kahler CW, Spillane NS, Leventhal AM, et al. Hostility and smoking cessation treatment outcome in heavy social drinkers. Psychol Addict Behav 2009;23:67–76.

[10] Hartmann-Boyce J, Chepkin SC, Ye W, et al. Nicotine replacement therapy versus control for smoking cessation. Cochrane Database Syst Rev 2018;5:CD000146. DOI: 10.1002/14651858.CD000146.pub5.

[11] Evins AE, Benowitz NL, West R, et al. Neuropsychiatric safety and efficacy of varenicline, bupropion, and nicotine patch in smokers with psychotic, anxiety, and mood disorders in the eagle's trial. J Clin Psychiatry 2019;39:108–16.

[12] Chang PY, Shiu MN, Yuan YT, et al. Comparative effectiveness of varenicline and nicotine replacement therapy for smoking cessation in older and younger smokers: a prospective cohort in Taiwan. Nicotine Tob Res 2017;21:149–55.

[13] Ebbert JO, Hughes JR, West RJ, et al. Effect of varenicline on smoking cessation through smoking reduction: a randomized clinical trial. JAMA 2015;313:687–94.

[14] Kotz D, Viechtbauer W, Simpson CR, et al. Cardiovascular and neuropsychiatric risks of varenicline and bupropion in smokers with chronic obstructive pulmonary disease. Thorax 2017;72:905–11.

[15] Yang M, Kiang M, Shang W. Filtering big data from social media-building an early warning system for adverse drug reactions. J Biomed Inform 2015;54:230–40.

[16] Jonnagaddala J, Liaw ST, Ray P, et al. Coronary artery disease risk assessment from unstructured electronic health records using text mining. J Biomed Inform 2015;58:S203–10.

[17] Lin FPY, Pokorny A, Teng C, et al. TEPAPA: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. Sci Rep 2017;7:6918https://doi.org/10.1038/s41598-017-07111-0.

[18] Ben Abacha A, Chowdhury MFM, Karanasiou A, et al. Text mining for pharmacovigilance: using machine learning for drug name recognition and drug-drug interaction extraction and classification. J Biomed Inform 2015;58:122–32.

[19] Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Association for Computational Linguistics, Ann Arbor, Michigan, U.S.A., 2005; 363–370.

[20] Yao J. Automated sentiment analysis of text data with NLTK. J Phys Conf Ser 2019;1187:052020.

[21] Wei CH, Kao HY, Lu Z. Pubtator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013;41:518–22.

[22] Inokuchi A, Washio T, Motoda H. An apriori-based algorithm for mining frequent substructures from graph data. European Conference on Principles of Data Mining and Knowledge Discovery Berlin, Heidelberg: Springer; 2000. 13–23.

[23] Wu X, Kumar V, Quinlan JR, et al. Top 10 algorithms in data mining. Knowl Inf Syst 2008;14:1–37.