Classifying Supplement Use Status in Clinical Notes

Yadan Fan, BS¹, Lu He, BS², Serguei V.S. Pakhomov, PhD¹,³, Genevieve B. Melton, MD, PhD¹,⁴, Rui Zhang, PhD¹,⁴
¹Institute for Health Informatics, ²Department of Computer Science, ³College of Pharmacy, ⁴Department of Surgery, University of Minnesota, Minneapolis, MN

Abstract

Clinical notes contain rich information about supplement use that is critical for detecting adverse interactions between supplements and prescribed medications. It is important to know the context in which supplements are mentioned in clinical notes to be able to correctly identify patients that either currently take the supplement or did so in the past. We applied text mining methods to automatically classify supplement use into four status categories: Continuing (C), Discontinued (D), Started (S), and Unclassified (U). We manually classified 1,300 sentences into these categories, which were further split as training (1000 sentences) and testing (300 sentences) sets. We evaluated the 7 types of feature sets and 5 algorithms, and the best model (SVM with unigram, bigram and indicator word within certain distance) performed F-measure of 0.906, 0.913, 0.914, 0.715 for status C, D, S, U, respectively on the testing set. This study demonstrates the feasibility of using text mining methods to classify supplement use status from clinical notes.

Introduction

The consumption of dietary supplements has risen dramatically over the last two decades worldwide. According to the World Health Organization (WHO), approximately 80% of the global population takes traditional herbal medicines as complementary or alternative medicine¹. Despite this widespread use, there remains a lack of information on the efficacy and safety of dietary supplements since the dietary supplements are classified and promoted as food by the United States DSHEA (Dietary Supplement Health and Education Act) of 1994². Supplements are often considered to be safe products and are taken with conventional medicines for the purpose of achieving a better healthcare outcome. However, concomitant use of prescribed medications and dietary supplements can lead to potentially dangerous adverse interactions, since supplements can affect the pharmacokinetic pathways of drugs. For example, warfarin can potentially interact with many supplements such as ginkgo, which can increase the risk of side effects like bleeding³.

Electronic health record (EHR) systems serve as the main healthcare documentation platform where supplement use information is documented. However, in some cases, structured data may not capture complete information about disease treatment, supplement use, etc⁴. A large amount of information about supplement use is embedded in clinical notes and can be further leveraged for clinical research and knowledge discovery. For example, a supplement may be mentioned in a context that indicates stopping its use, or in a context in which the use of the supplement is only being considered or discussed. In order to support clinical research on supplement safety, it is important to extract the information of supplement use, such as the history (e.g., past use) and current status (e.g., active or discontinued).

Methods to identify medication use status from clinical notes have been investigated in previous studies⁵,⁶. In two of these studies, rule-based and machine learning algorithms (e.g., Support Vector Machine (SVM), Maximum Entropy
(ME)) along with indication features were used to detect medication status from free text\(^5\). Another study relied on keywords to establish supplement use without further analysis of the supplement status\(^7\). To the best of our knowledge, classification of supplement usage status in clinical notes through text mining methods has not been extensively investigated.

The objective of this study was to examine the performance of several text mining techniques for automatic categorization of the status context in which supplements are found in unstructured text of clinical notes.

**Materials and Methods**

In this study, we used several text mining methods to automatically classify the status of supplement mentions in clinical notes into four categories (Continuing, Discontinued, Started, and Unclassified). In this study, we focused on 25 most popular supplements available in US pharmacies, which include alfalfa, echinacea, fish oil, garlic, ginger, ginkgo, ginseng, melatonin, St John’s Wort, Vitamin E, bilberry, biotin, black cohosh, coenzyme Q10, cranberry, dandelion, flax seed, folic acid, glucosamine, glutamine, kava, lecithin, milk thistle, saw palmetto, and turmeric. Seven types of feature sets were trained using 5 classification algorithms on the 1,000 sentences containing 10 supplements. To ensure the validity of our method, we tested the most optimal model on 300 sentences containing another 15 supplements.

**Data Collection and Gold Standard**

To collect notes containing the above-mentioned 25 supplements, we first searched for the notes using the supplement names and their corresponding variations (Table 1). Disease related supplement mentions were filtered out, such as “Vitamin E deficiency” or “Vitamin E level in the blood”. To build up the expert-curated gold standard, we randomly selected 1,300 sentences containing 25 most commonly consumed supplements (Table 1) complied from 75,000 clinical notes retrieved from the University of Minnesota’s Clinical Data Repository (CDR).

In order to generate a reference standard, a total of 1,300 sentences were independently annotated by two annotators. Firstly, we randomly selected 100 sentences to build the annotation guidelines. Disagreements were resolved with discussion to reach consensus. We evaluated inter-rater agreement based on these 100 sentences using Cohen’s kappa and percentage agreement. The remaining 1,200 sentences were equally split into two parts, which were independently classified by two annotators. For each supplement mention, the annotator assigned it to one of the four classes according to the contextual information: Continuing (C), Discontinued (D), Started (S), and Unclassified (U). Specifically, status C indicates that there is evidence showing the patients continue on the current supplements (e.g., “Decrease fish oil to 1 tablet daily”). Status D indicates the discontinuation of supplements for certain reasons like the allergic reaction, ineffectiveness or potential adverse interactions with prescribed medications (e.g., “Stopped taking her garlic tablets a week ago”). Status S refers to the initiation of new supplements or restarting supplements (e.g., “Begin melatonin 10mg 1 hour before bedtime”). Status U is associated with mentions of supplements that do not offer ample information about the use status, such as recommendation or patient’s education (e.g., “Recommended cranberry pills”).

The 1,300 annotated sentences were then split into two parts. The first part is the training set which consists of 1,000 sentences (~77% of the gold standard) of 10 supplements including alfalfa, echinacea, fish oil, garlic, ginger, ginkgo, ginseng, melatonin, St John’s Wort, and Vitamin E. 100 sentences or statements (incomplete sentences) were randomly selected for each of the 10 supplements. We used the training set to select the best algorithm with the
feature sets. The second part is the 300 sentences (~23% of the gold standard) of the remaining 15 supplements, including bilberry, biotin, black cohosh, coenzyme Q10, cranberry, dandelion, flax seed, folic acid, glucosamine, glutamine, kava, lecithin, milk thistle, saw palmetto, and turmeric. 20 sentences were randomly selected for each of the 15 supplements. The held-out testing data set was used to evaluate the best performing model developed using cross-validation on the training set.

Table 1. Supplement name and lexical variations.

| Supplement          | Lexical variations                      |
|---------------------|----------------------------------------|
| alfalfa             | alfalfa                                |
| echinacea           | echinacea                              |
| fish oil            | fish oil                               |
| garlic              | garlic                                 |
| ginger              | ginger                                 |
| ginkgo              | ginkgo, ginko, gingko, ginkoba         |
| ginseng             | ginseng                                |
| melatonin           | melatonin                              |
| St John’s Wort      | St. John’s Wort, St. Johns Wort, St. John Wort, St John Wort |
| Vitamin E           | Vitamin E, Vit E                       |
| bilberry            | bilberry                               |
| biotin              | biotin                                 |
| black cohosh        | black cohosh                           |
| Coenzyme Q10        | Coenzyme Q 10, Coenzyme Q-10, CoQ10    |
| cranberry           | cranberry                              |
| dandelion           | dandelion                              |
| flax seed           | flax seed, flaxseed                    |
| folic acid          | folic acid                             |
| glucosamine         | glucosamine                            |
| glutamine           | glutamine                              |
| kava                | kava                                   |
| lecithin            | lecithin                               |
| milk thistle        | milk thistle                           |
| saw palmetto        | saw palmetto                           |
| turmeric            | turmeric                               |

Model and Feature Selection

We trained and evaluated 5 machine learning algorithms, including Support Vector Machine (SVM), Maximum Entropy, Naïve Bayes, J48, and Random Forest on 7 different sets of features using Weka.

Pre-processing: before training, we cleaned the sentences by removing numbers, punctuation, and special characters
such as “*”, Stop-words were also removed. Typical words “on”, “off” and some negated words like “no”, “not” were kept, since some sentences such as “He is on fish oil”, “Pt says he has been off Black Cohosh for about 2 months” and “No supplements and alfalfa” indicate the status of supplement use.

Feature sets: The 7 different types of features are described below as follows:

1) Type 0 - raw unigrams
   The feature set applied the bag-of-words representation method.

2) Type 1 - normalized unigrams
   We applied lexical variation generation (LVG) tool to normalize different lexical variations of the same term to one form to reduce the dimensions of feature sets. For example, tokens of “take”, “taking”, “takes”, “taken”, and “took” were normalized as “take.” Each distinct token corresponds to a feature vector that was recorded in a binary occurrence matrix. Type 1 only included the unigrams of each sentence.

3) Type 2 – normalized unigrams + bigrams
   Type 2 model included both normalized unigrams used in Type 1 and bigrams were also added.

4) Type 3 – indicator words only
   A series of lexical cues can be used as status indicators. For example, “She has increased her alfalfa tabs and this has eliminated her symptoms of chest tightness” may indicate status C, “Pt reports that she has discontinued her Vitamin E” and “Stopped her aspirin and fish oil” may indicate status D, “Pt started taking ginkgo biloba” may indicate status S, and “Melatonin is recommended for sleep aid” and “Pt states that someone else suggested taking Vitamin E” may indicate status U. From the instances above, “increased”, “discontinued”, “stopped”, “started”, “suggested”, and “recommended” are all indicators of use status. A list of such indicator words was constructed by empirical observation, which was shown in Table 2. Our Type 3 model only included these indicator words, which were arranged in a feature vector with a binary value indicating the presence or absence in the context of the supplement mention.

5) Type 4 – normalized unigrams + indicator words with distance
   The indicators of status are usually found close to the supplement mentions, thus it’s significant to set the distance between the occurrence of the indicator words with respect to the supplements mention, which means the number of tokens between them. If the indication word is too far away from the supplement, it might not indicate the use status of the supplements. For example, “he continues on Coumadin and also has recently started ginseng as he is concerned about the fatigue he will have during chemotherapy,” contains two indication words: “continue” and “start”. In order to decide upon the optimal window size, we started by taking 1 token on each side of the supplement text, and increased to 5 tokens on both sides of the supplement text (left and right). If the supplement is at the beginning of a sentence, only tokens to the right of the supplement name were retrieved. The weighted F-measure was calculated every time the distance increased by one token. We chose the optimal distance by the value of F-measure. The Type 4 model included the predefined lexical cues listed in Table 2 as binary features within a certain distance to the supplements mention along with normalized unigrams.

6) Type 5 – normalized unigrams + bigrams + indicator words with distance
Type 5 included normalized unigrams, bigrams, indication words within certain distance. The distance was the optimal value in Type 4.

7) Type 6 – nouns + verbs + adverbs

Based on our observations, we found that verbs may hold more valuable information about the use status, which was reflected by the indication words listed in Table 2. To verify our observation, only verbs, nouns and some of the adverbs such as “no,” “not,” and “never” were sorted out for each sentence. Stanford Parser\textsuperscript{10} was used to identify the POS tags of each word. We only focused on the nouns (NN/NNS/NNP/NNPS), present tense verbs (VB/VBG/VBP/VBZ), past tense verbs (VBD/VBN) and some of the adverbs (RB) such as “no,” “not,” “never.” Type 6 model only incorporated nouns, verbs, and some of the adverbs.

Table 2. Indicator words and their corresponding morphological forms.

| Indication Words | Keywords |
|------------------|----------|
| start            | start, starts, started, starting |
| begin            | begin, begins, began, begun |
| restart          | restart, restarts, restarted, restarting |
| resume           | resume, resumes, resumed, resuming |
| initiate         | initiate, initiates, initiated, initiating |
| add              | add, adds, added, adding |
| try              | try, tries, tried, trying |
| increase         | increase, increases, increased, increasing |
| decrease         | decrease, decreases, decreased, decreasing |
| reduce           | reduce, reduces, reduced, reducing |
| lower            | lower, lowers, lowered, lowering |
| continue         | continue, continues, continued, continuing, continuation |
| take             | take, takes, took, taking, taken |
| consume          | consume, consumes, consumed, consuming |
| tolerate         | tolerate, tolerates, tolerated, tolerating |
| stop             | stop, stops, stopped, stopping |
| discontinue      | discontinue, discontinues, discontinued, discontinuing |
| hold             | hold, holds, held, holding |
| recommend        | recommend, recommends, recommended, recommendation |
| advise           | advise, advises, advised, advising |
| avoid            | avoid, avoids, avoided, avoiding |
| deny             | deny, denies, denied, denying |
| decline          | decline, declines, declined, declining |
| refuse           | refuse, refuses, refused, refusing, refusal |
| neg              | no, not, never |
A 10-fold cross validation was performed for each experiment. The precision, recall, and F-measure were used as the measurement. The model with the best performance was then applied on the held-out testing set, and the precision, recall, and F-measure were also calculated.

Results

Comparing manual classifications on the gold standard resulted in a Cohen’s Kappa of 0.93 and the percentage agreement of 95%, which indicates that our gold standard can be considered reliable and useful to build models and evaluate classification results. A total of 1,000 annotated sentences and statements in the training data set consisted of 380 with Status C, 156 with Status D, 139 with Status S, and 325 with Status U. Among the 300 clinical notes in the testing data set, there were 112 sentences for Status C, 47 for Status D, 74 for Status S, and 67 for Status U.

The performance of our classifiers on Type 4 model with the distance from the indication words to both sides of the supplements mentions ranging from 1 to 5 tokens were shown in Table 3. SVM outperformed than the other four classifiers even though its performance is close to that of Random Forest. The weighted F-measure increased with the distance until it reached its peak; when the distance was 4 or 5 tokens from the supplement text (weighted averaged F-measure: 0.798). We arbitrarily selected the window size of 4 tokens on both sides of the supplement text (L4 R4).

| Distance from the supplement mentions | SVM  | Maximum Entropy | Naïve Bayes | Decision Tree | Random Forest |
|---------------------------------------|------|-----------------|-------------|---------------|---------------|
| L1 R1*                                | 0.630| 0.625           | 0.462       | 0.582         | 0.605         |
| L2 R2                                 | 0.741| 0.699           | 0.472       | 0.683         | 0.740         |
| L3 R3                                 | 0.780| 0.758           | 0.526       | 0.736         | 0.783         |
| L4 R4                                 | 0.798| 0.751           | 0.541       | 0.745         | 0.786         |
| L5 R5                                 | 0.798| 0.776           | 0.545       | 0.740         | 0.796         |

* L1R1 indicates context from the 1 word on the left side to the 1 word on the right side of the supplement mention.

Performance of the 5 classifiers on 7 sets of features in the training set is shown in Table 4. In general, for each type of feature set, SVM outperformed other classifiers. Using the SVM classifier, it seemed that the performances of all the 7 models were enhanced compared with the baseline. Specifically, the Type 5 model had the best performance with the highest F-measure of 0.844. For the maximum entropy classifier, the Type 5 model also achieved the highest F-measure of 0.800 compared with other models. However, Naïve Bayes seemed to have really poor performance on the 7 models, with the best F-measure at the baseline. For the decision tree, the Type 3 model had the best performance with the F-measure of 0.788. Interestingly, unlike other four classifiers, random forest reached its highest F-measure (0.812) with Type 3 and Type 6 models. Overall, the SVM classifier outperformed other four classifiers in the training data, with F-measure ranging from 0.748 of baseline to 0.844 of Type 5 model. Naïve Bayes had the poorest performance, with the lowest F-measure of 0.497 and the highest F-measure of 0.721.
Table 4. The performance of 5 classifiers on 7 types of feature sets on training set.

| Feature Sets | Classifiers          |
|--------------|----------------------|
|              | SVM                  | Maximum Entropy | Naïve Bayes | Decision Tree | Random Forest |
|              | P  R  F              | P  R  F        | P  R  F    | P  R  F      | P  R  F      |
| Type 0       | 0.771 0.751 0.748    | 0.778 0.762 0.760 | **0.757** 0.726 **0.721** | 0.738 0.718 0.717 | 0.789 0.763 0.753 |
| Type 1       | 0.799 0.772 0.760    | 0.735 0.734 0.734 | 0.659 0.639 0.596 | 0.792 0.759 0.743 | 0.791 0.767 0.756 |
| Type 2       | 0.839 0.838 0.838    | 0.813 0.794 0.786 | 0.635 0.579 0.497 | **0.804** 0.790 0.785 | 0.772 0.753 0.747 |
| Type 3       | 0.789 0.790 0.788    | 0.784 0.783 0.783 | 0.750 **0.729** 0.711 | 0.792 **0.792** **0.788** | **0.818** **0.815** **0.812** |
| Type 4       | 0.799 0.798 0.798    | 0.793 0.761 0.751 | 0.678 0.612 0.541 | 0.761 0.75 0.745 | 0.816 0.794 0.786 |
| Type 5       | **0.845** 0.845 0.844 | **0.823** 0.806 **0.800** | 0.653 0.584 0.499 | 0.792 0.778 0.772 | **0.818** 0.810 0.808 |
| Type 6       | 0.829 0.829 0.828    | 0.750 0.749 0.749 | 0.681 0.647 0.613 | 0.791 0.787 0.784 | **0.818** 0.813 0.812 |

The performance of the SVM classifier on the training data and testing data in terms of 4 status categories are shown in Table 5 and Table 6, respectively. In the training set, status D had the highest precision (0.951) and F-measure (0.913) compared with other statuses. Status C had the largest recall of 0.895. In comparison, performances for all statuses in the testing set increased, with the exception of Status U. F-measures for Statuses C, D, and S were all over 0.900. Highest recall and precision was 0.946 for status C and 0.933 for status D.

Table 5. The performance of SVM classifier with type 5 feature set on the training data in 4 status categories.

| Category           | Precision | Recall | F-measure |
|--------------------|-----------|--------|-----------|
| Continuing (C)     | 0.835     | **0.895** | 0.864     |
| Discontinued (D)   | **0.951** | 0.878  | **0.913** |
| Started (S)        | 0.842     | 0.842  | 0.842     |
| Unclassified (U)   | 0.806     | 0.769  | 0.787     |

Table 6. The performance of SVM classifier with type 5 feature set on the testing data in 4 status categories.

| Category           | Precision | Recall | F-measure |
|--------------------|-----------|--------|-----------|
| Continuing (C)     | 0.869     | **0.946** | 0.906     |
| Discontinued (D)   | **0.933** | 0.894  | 0.913     |
| Started (S)        | 0.896     | 0.932  | **0.914** |
| Unclassified (U)   | 0.786     | 0.657  | 0.715     |

Among the 7 types of models with different feature sets, the Type 5 model had the best performance with the highest F-measure of 0.844. It is also evident from the results in Table 5 that the Type 5 SVM model performed very well on predicting most of the categories. It achieved the highest F-measure score in the Discontinued (D) category (F-measure: 0.913) and the lowest in the Unclassified (U) category (F-measure: 0.787).

Due to the high performance of Type 5 SVM model in the training data, we chose this model to perform final validation on the held-out set of 300 clinical notes with 15 supplements. According to the results shown in Table 6,
we can see that the SVM classifier trained on a set of 10 supplements can accurately identify the use status in the testing set containing completely different 15 supplements. This is especially true for the Started (S) category, with the F-measure reaching 0.914. It also performed moderately well on the Unclassified (U) category, with the F-measure of 0.715.

Discussion

A fair amount of information about supplement use is stored in clinical notes. Identifying the use status, especially the active status (e.g., started, continuing) or use history (e.g., discontinued), is critical for clinical research and data analysis, such as detecting the adverse interactions between prescribed medications and dietary supplements or determining the efficacy of certain supplements. Extracting such information would be the first step to support clinical research for supplements.

In this study, SVM outperformed other algorithms. Various feature sets were evaluated by 5 popular algorithms. Lexical normalization technique is widely used for information retrieval and information extraction. It seems useful to improve the performance after comparing the Type 1 to baseline model, which did not use the normalization technique. Introducing bigrams to the feature set significantly improved the performance. Compared to the unigram, bigrams provide some contextual information to the model. Distance is another parameter we considered to optimize the model. Distance or window size can control the size of the context of supplements mention. Results show that only certain distances (4 tokens in this case) can reach the best performance. Incorporating the indication words within certain distance of supplement mention can somewhat improve the performance of the classifiers. We have not tested on a larger window size, as many statements are too short for this to be a useful approach. The most optimal model in our study has combined a variety of features: modified unigrams, bigrams, indication words within certain distance (4 tokens on both sides of the supplement mention). Besides, the performance of the Type 6 model confirmed our observation that verbs play a significant role in the classification task. However, Stanford parser was originally designed for clinical free texts, and we found that it cannot accurately identify POS tags for statements without complete syntactic structures. This limitation also made the performance of Type 6 model insufficient for this task. Exploring other parsers may improve the performance. For example, Trigrams’n’Tags (TnT), a hidden Markov model, has shown a promising performance on a set of synthetic clinical notes.

Our study also has several limitations. First, the data set is relatively small. In order to enhance the performance of the model, more training data is needed. More indication words will be included. We also found the typos and abbreviations by clinicians also make this task challenging. Clinicians will refer to “Discontinue” as ‘DC,’ refer “discontinued” as “d/ced,” or “de\'d.” Failure to recognize these abbreviations will also impede the performance of the classification work. Incorporating existing abbreviation disambiguation methods could potentially enhance the performance. In future studies, exploring the temporal information may also help to refine the absolute start and discontinue dates for supplements.

Conclusion

We applied text mining methods to automatically identify the use status of supplements from clinical notes. The results demonstrated that our model built from 10 supplements performs well in predicting the use status of another 15 supplements, with F-measure ranging from 0.906 to 0.914 for Statuses C, D, and S, demonstrating good generalizability. This study shows that applying text mining methods on clinical notes can successfully extract the detailed and rich information about supplement use. The extracted information can be further applied to detect
adverse events and interactions between supplements and drugs in clinical settings.

Acknowledgement

The study was partly supported by the University of Minnesota Grant-in-aid award (Zhang) the Agency for Healthcare Research & Quality grant (#1R01HS022085) (Melton), and the University of Minnesota Clinical and Translational Science Award (#8UL1TR000114) (Blazer). The clinical data was provided by the University of Minnesota’s Clinical Translational Science Institute (CTSI) Informatics Consulting Service. The authors thank Reed McEwan and Fairview Health Services for their data support of this research.

Reference

1. Choi JG, Eom SM, Kim J, Kim SH, Huh E, Kim H, Lee Y, Lee H, Oh MS. A Comprehensive Review of Recent Studies on Herb-Drug Interaction: A Focus on Pharmacodynamic Interaction. The Journal of Alternative and Complementary Medicine. 2016 Apr 1;22(4):262-79.
2. Bailey RL, Gahche JJ, Lentino CV, Dwyer JT, Engel JS, Thomas PR, Betz JM, Sempos CT, Picciano MF. Dietary supplement use in the United States, 2003–2006. The Journal of nutrition. 2010 Dec 1:jn-110.
3. Hu Z, Yang X, Ho PC, Chan SY, Heng PW, Chan E, Duan W, Koh HL, Zhou S. Herb-drug interactions. Drugs. 2005 Jun 1;65(9):1239-82.
4. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012 Jun 1;13(6):395-405.
5. Sohn S, Murphy SP, Masanz JJ, Kocher J, Savova GK. Classification of medication status change in clinical narratives. InAMIA Annu Symp Proc 2010 Nov 13 (Vol. 2010, pp. 762-6).
6. Pakhomov SV, Ruggieri A, Chute CG. Maximum entropy modeling for mining patient medication status from free text. InProceedings of the AMIA Symposium 2002 (p. 587). American Medical Informatics Association.
7. Stoddard GJ, Archer M, Shane-McWhorter L, Bray BE, Redd DF, Proulx J, Zeng-Treitler Q. Ginkgo and Warfarin Interaction in a Large Veterans Administration Population. InAMIA Annual Symposium Proceedings 2015 (Vol. 2015, p. 1174). American Medical Informatics Association.
8. Weka. http://www.cs.waikato.ac.nz/ml/weka/
9. LVG. https://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_004.html
10. The Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml
11. Knoll BC, Melton GB, Liu H, Xu H, Pakhomov SV. Using synthetic clinical data to train an HMM-based POS tagger. In2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) 2016 Feb 24 (pp. 252-255). IEEE.