HHS Public Access

Author manuscript

Stud Health Technol Inform. Author manuscript; available in PMC 2018 January 09.

Published in final edited form as:

Stud Health Technol Inform. 2017 ; 245: 614–618.

Mining Adverse Events of Dietary Supplements from Product Labels by Topic Modeling

Yefeng Wang[a], **Divya R. Gunashekar**[b], **Terrence J. Adam**[a,c], and **Rui Zhang**[a,d]

[a]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA

[b]School of Public Health, University of Minnesota, Minneapolis, MN, USA

[c]College of Pharmacy, University of Minnesota, Minneapolis, MN, USA

[d]Department of Surgery, University of Minnesota, Minneapolis, MN, USA

Abstract

The adverse events of the dietary supplements should be subject to scrutiny due to their growing clinical application and consumption among U.S. adults. An effective method for mining and grouping the adverse events of the dietary supplements is to evaluate product labeling for the rapidly increasing number of new products available in the market. In this study, the adverse events information was extracted from the product labels stored in the Dietary Supplement Label Database (DSLD) and analyzed by topic modeling techniques, specifically Latent Dirichlet Allocation (LDA). Among the 50 topics generated by LDA, eight topics were manually evaluated, with topic relatedness ranging from 58.8% to 100% on the product level, and 57.1% to 100% on the ingredient level. Five out of these eight topics were coherent groupings of the dietary supplements based on their adverse events. The results demonstrated that LDA is able to group supplements with similar adverse events based on the dietary supplement labels. Such information can be potentially used by consumers to more safely use dietary supplements.

Keywords

Dietary Supplements; Natural Language Processing; Pharmacovigilance

Introduction

According to a recent cross-sectional study using the data from the National Health and Nutrition Examination Survey (NHANES) over the time period from 199–2012, 52% of U.S. adults use dietary supplements [1]. Americans spend more than $28 billion annually on dietary supplements in the belief that they are able to improve health and are generally harmless in their side effects [2]. However, a study conducted by the Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC) has shown

Address for correspondence: Rui Zhang, PhD, University of Minnesota, Minneapolis, MN, USA zhan1386@umn.edu.

that dietary supplements caused on average about 23,000 emergency room visits every year [3].

The distinct regulatory framework for dietary supplements is one of the major reasons why dietary supplement safety requires scrutiny. In 1994, the Dietary Supplement Health and Education Act (DSHEA) was passed. According to DSHEA, dietary supplements belong to a subcategory of food, and can be waived from the premarket approval for efficacy and safety testing that are required for new drugs and food additives [4, 5]. While the DSHEA did give the consumers a great variety of choices in dietary supplements, it also exposed consumers to considerable risks in terms of dietary supplement safety [6].

A major clinical concern is that the adverse events of the active ingredients in the supplements may be unknown. For example, the sexual enhancement supplement Zotrex, which had been recalled by FDA in 2011, contained sulfoaildenafil, an ingredient whose adverse effects on humans were never tested. It is estimated that more than 50,000 new dietary supplements have been available on the market since the DSHEA became law [2]. DSHEA implementation addressed dietary supplement safety concerns by requiring manufacturers to provide adequate proof to the FDA that the new ingredients introduced are safe [4, 5]. The regulation was poorly enforced, as the FDA received appropriate notifications for only 170 new ingredients [2] potentially making consumers more vulnerable to these potential supplement adverse events.

Literature review has been used to study the adverse events of dietary supplements. Pittler, Schmidt, and Ernst conducted a systematic review on the adverse events of herbal food supplements for body weight reduction including 33 case reports and 48 clinical trials dating from 1995 to 2004 [7]. Seven ingredients were covered in this review and their adverse events were reported in detail. However, literature-based studies suffer from several drawbacks. Firstly, literature-based reviews can only cover a few dietary supplements. In the example above, the authors could only focus on seven ingredients intended for body weight reduction, which is only a tiny portion of the dietary supplement ingredients currently available on the market. Secondly, the qualities of the literature are inconsistent. The literature that reported the adverse events did not explain the related mechanism or establish any cause and effect relationship. Finally, incompleteness of the literature is an intrinsic limitation of literature-based studies [7, 8].

Alternatively, dietary supplement product labels contain vital information on the adverse events of dietary supplements. The FDA has stringent rules on dietary supplement labeling information including statement of identity, active ingredients, manufacturer information, structure or function claims, and health claims to make sure the information is truthful and not misleading [4,5,9]. Consequently, the supplement labels provide an opportunity for more efficient methods of supplement adverse events related research. However, unlike drug labeling, adverse events related to the supplements typically exist in the precaution section of labels in free text format. Thus, text mining techniques are required to extract adverse events information from supplement labels.

Topic modeling is a set of text mining methods that automatically discover the underlying themes of a collection of documents without prior document annotation or labeling. Implementations of topic modeling includes latent semantic indexing (LSI), probabilistic latent semantic analysis (PLSA) [10], and latent Dirichlet Allocation (LDA). LDA is one of the more common topic model techniques in the current literature. A topic is defined as a "distribution over a fixed vocabulary" [11]. LDA is a statistical model, which assumes that all the documents in the collection can be described by a group of topics, but each document is a different distribution of these topics. As a generative graphic model, LDA can discover the underlying topic distribution for a large document collection.

Topic modeling has been applied to social media and drug product labeling to discover new knowledge. Sullivan et al. have recently applied LDA to Amazon.com nutritional supplement reviews to find potentially unsafe dietary supplements [12]. Bisgin *et al.* have used LDA on FDA drug labels to group drugs with similar safety concerns and therapeutic uses [13], and this information was later used to discover drug repositioning opportunities [14]. However, the methodology has to the best of our knowledge not been applied to dietary supplement labels. Product labels provide useful information to group dietary supplements based on the listed adverse events. Thus, in this study, we tried to demonstrate that the application of the LDA technique on the precautions statements of dietary supplement labels can yield useful groupings of adverse events without any prior knowledge to analyze the adverse events.

Dietary Supplement Label Database (DSLD)

The DSLD is created and managed by Office of Dietary Supplement (ODS) and U.S. National Library of Medicine (NLM) in the National Institutes of Health (NIH) [15]. The database includes complete label contents among 55,456 dietary supplement products currently marketed, off the market, and consumed by NHANES participants. For each supplement product, there are four distinct label sections: namely, product information (including product name, statement of identity, serving information, and target groups); dietary supplement facts (including active ingredients); label statements (including formulation, precaution, and suggested use); and contact information (including manufacturer information).

DSLD provides a web application programming interface (API) for an efficient extraction of the labeling information in JavaScript Object Notation (JSON) format. It also provides a comma-separated value (CSV) file including database identifiers (DSLD ID) for all the products in the database.

Side Effect Resource (SIDER)

The SIDER database is a free resource that contains 5,868 adverse drugs reactions (ADRs) of 1430 drugs. It uses the Medical Dictionary for Regulatory Activities (MedDRA) terms to describe the drug side effects. The list of side effect MedDRA terms was extracted from the drug labels and downloaded in CSV using SIDER version 4.1 [16].

Although the SIDER database is not designed for dietary supplements, the adverse events encountered with the dietary supplements are similar to those with drugs. Therefore, the

SIDER MedDRA terms list was used to recognize adverse event keywords in the DSLD labels.

Methods

Figure 1 illustrates the overview of the methods. We extracted the supplement label information from the DSLD, preprocessed the label statements to create a collection of supplement documents represented by a list of SIDER MedDRA terms, and then grouped supplements using LDA. The adverse events topics generated by LDA were further evaluated by human annotation.

DSLD Label Preprocessing

The labels of 55,456 dietary supplements available in DSLD were extracted via the API in JSON format. Two criteria were applied to generate a subset of dietary supplements products for topic modeling analysis.

1. Products that only have one active ingredient entry in the dietary supplement facts section were selected for further preprocessing to prevent interactions between ingredients from confounding our analysis.
2. Products that have one or more statements under the “Precautions” subsection of the label statements section.

The “Precautions” statements of each supplement in the subset were split into a list of single sentences with the punctuation marks removed and all words lowercased. We mapped terms (allowing a window size of 5 for multiple-token terms) to the MedDRA terms listed in SIDER for each supplement product. A windowed mapping was used since a multiple-token phrase may be expressed noncontiguously or in different order. For example, the following statement, “Beta-alanine may cause a harmless, temporary tingling or flushing sensation”, should match the word “tingling sensation” in the MedDRA terms list although the two words do not appear next to each other. A windowed mapping will yield the correct result, while an exact match will not.

It is noteworthy that although the supplements in the dataset all have only one active ingredient entry, some of the ingredients are actually blends. For example, “chocamine” was registered as a single ingredient, but in fact it consists of a mixture including theobromine, caffeine, theophylline, phenylethylamine, tyramine, phenylalanine, tryptophan, and tyrosine. This type of supplements was excluded in our analysis due to the complexity of the components in blends.

The windowed mapping generates a “document”, which is a list of the MedDRA terms appeared in the “Precautions” statement for each supplement. In other words, each supplement is represented by a list of MedDRA terms. The collection of the documents in the subset was then analyzed by LDA model.

Topic Modeling Analysis

The LDA model was implemented using the Python *gensim* package [17]. Fifty topics were generated by the training process using the dataset of preprocessed labels. To simplify the analysis, we assigned each supplement only one topic. Each topic was represented by the topic keywords, i.e. MedDRA terms, and their corresponding probabilities. As discussed in the introduction, a topic is a distribution over all the MedDRA terms which appeared in the label statements. Therefore, to better define a topic in terms of adverse events, the topic keywords with probabilities lower than 0.05 were not considered to define a topic. Due to the probabilistic nature of the LDA model, the training was repeated 10 times to see if there are significant differences among the generated topics.

The conditional probability of each topic given a particular supplement was calculated, and the supplement was assigned to the topic that has the maximal conditional probability. The distribution of the supplements over the topics was then derived by counting the number of supplements under each topic.

Topic Evaluation

Since the topics generated by LDA are not based on any prior annotation or labeling of the statements, it is necessary to evaluate the topics by checking if the supplements assigned to a certain topic are actually related to the adverse events the topic represents. Two criteria were used to select the topics to be evaluated:

1. The chosen topics all have more than one keyword with probabilities greater than 0.05. This study is focused on grouping of dietary supplements by their adverse events. If a topic is defined by more than one adverse event, then it is necessary to examine if these adverse events are related in terms of supplement function.
2. The number of supplements under the chosen topics should be within the range of 10–50. If there are too few products under a certain topic, there would be insufficient data points to generate any significant grouping patterns; on the other hand, if there are too many products under a certain topic, the grouping may not be meaningful, as a group may contain a mixture of completely different supplements.

Eight topics out of 50 were chosen accordingly for evaluation based on the group of adverse events indicated by the topic keywords. The largest topic contained 46 products, while the smallest topic contained 13 products.

Two human annotators manually reviewed the active ingredients and the health claims in the statements on each supplement labels under the eight topics. Annotators made a consensus on their annotations when there was a disagreement on a product assigned for a given topic. To quantitatively evaluate the performance of topic modeling, we used the metric called topic relatedness, defined as the percentage of products (or ingredients) related to the topic keywords out of the total number of products (or ingredients) assigned to the same topic.

Results

DSLD Label Preprocessing

There are 6,123 side effect terms that were extracted from the drug labels in the MedDRA term list. Among 55,456 dietary supplements available in DSLD, 27.9% (15,452) of the supplement products have only one active ingredient entry. Only 3.6% (2,014) of these single ingredient supplements whose “Precautions” label statements contained adverse events were matched to a total of 3.9% (239) terms within the MedDRA term list.

Topic Modeling Analysis

The distribution of the 2,014 supplements over the 50 topics generated by the LDA was shown in Figure 2. The topic with the most supplements has 234 products, while the topic with the least supplements has only 1 product. The median of the supplement counts over the 50 topics is 21, and the first quartile and the third quartile is 8 and 39, respectively.

The top three topics in supplement counts grouped 663 (32.9%) supplements that have precautions related to pregnancy and breast feeding, and all of these three topics have only a single topic keyword to represent their adverse events. However, these topics did not satisfy the criteria for further evaluation. Therefore, another eight topics including 210 products were selected for evaluation as shown in Table 1. The number of one-ingredient products and the unique ingredients of each topic was shown in Table 2.

Topic Evaluation

The topic relatedness of each topic on the product and the ingredient level was shown in Figure 3. Topic 24, 29, and 47 have reached a topic relatedness of 100% at both product level and ingredient level. This suggests that all the supplements grouped under these topics have mentioned the adverse events given in the topic keywords in their label statements.

The topic relatedness of the remaining five topics indicates that there exist supplements that do not contain statements relevant to the adverse events mentioned in the corresponding topic keywords. The discrepancies between the product and the ingredient level relatedness are also different among the topics. For Topic 1, 33, and 42, the product level topic relatedness is higher than the ingredient level one. The opposite is true for Topic 11 and 12. Topic 12 had the lowest product level topic relatedness at 58.8%, and Topic 33 had the lowest ingredient level topic relatedness at 57.1%.

The inconsistent differences for topic relatedness between product level and ingredient level relatedness among different groups are due to the fact that some products in the same group share the same ingredients. If there exist several products containing the same ingredient that are related to the adverse events in the topic, a lower topic relatedness at ingredient level is expected. For example, there are 22 products that contain vitamin D3 under Topic 1 and vitamin D3 was deemed a related ingredient that may cause these adverse events. Therefore, a lower ingredient level relatedness for this topic was obtained.

Conversely, if there exist several products containing the same ingredient that are irrelevant to the adverse events in the topic, a higher ingredient level topic relatedness will be

observed. The opposite example could be found in the topic related to obstruction, inflammation, intestinal obstruction, pain, and abdominal pain. Four products containing celandine were grouped under Topic 12, but celandine is an ingredient that was judged as not related to the adverse events mentioned in the topic. As a result, a higher ingredient level topic relatedness was observed.

Discussion

The adverse effects of dietary supplements were brought under more careful scrutiny due to the rapid growth of the supplement market and consumer base. The special regulatory framework for dietary supplements has led to both challenges and opportunities for research in supplement adverse events. The challenge lies in the fact that the soaring number of new supplement products introduced to the market demands more efficient ways to study the adverse events of a wide variety of dietary ingredients, while the opportunity arises in the supplement labels where relevant information of adverse events was provided as strictly required by the FDA. Topic modeling, specifically LDA, is a powerful method to find hidden grouping patterns within a series of documents without any prior knowledge or annotation. Therefore, the objective of this study is to utilize LDA as an effective method to generate useful groupings of dietary supplements based on the adverse events found in their supplement labels.

Since each topic generated by LDA is represented by a list of adverse events, we evaluated a subset of these topics by human annotation whether the supplements under each topic were actually related to the adverse events indicated by the topic keywords. If a topic is represented by multiple adverse events, the relationship among them was further examined to determine if there are multiple sub-topics.

Our evaluation results have shown that LDA was able to find similar supplement products and categorize them into the same topic based only on the adverse events mentioned in product labels. These supplements may contain the same active dietary ingredient or different ingredients with similar functions. For instance, in Topic 11, four different products with the same ingredients SAMe and six products containing dietary ingredients that are all different derivatives of androsterone were grouped together due to their potential risk of causing anxiety and depression.

We also found that LDA could associate clinically related adverse events with each other within the same topic. Among the eight topics listed in Table 1, five were highlighted in bold because the adverse events representing each topic were related and forming a single coherent topic. For example, in Topic 1, hyperparathyroidism, lymphoma, and sarcoidosis are able to cause hypercalcemia [18]. Bipolar disorder patients who are taking lithium medication are at risk of lithium-induced hypercalcemia [19]. This suggests that the dietary supplements reviewed under this topic were all linked to calcium metabolism. Although hypercalcemia did not appear in the dietary supplement labels, it is the hidden commonalities among the four adverse events that represent the topic. The most frequent ingredients that appeared under this topic is vitamin D3, which is an important factor in the calcium metabolism [20]. This suggests that consumers with any one of the pre-existing

conditions mentioned in the topic may expose themselves to adverse events by taking vitamin D3. Thus, the above examples demonstrated that LDA had successfully grouped dietary supplements based only on their possible adverse effects listed in the product labels into a coherent topic. The topics can give insight into further literature-based studies on the adverse effects of a particular set of dietary supplements.

The study has its limitations. Firstly, the MedDRA terms included in the SIDER does not cover all possible synonyms that may appear in the label statements. For example, loss of hair or hair loss, which are the synonyms for alopecia, are not included in the MedDRA terms list, but were present in supplement labels. The windowed mapping method requires every word in the MedDRA terms to be present in the statements, and therefore was not able to capture all the variations of the same adverse effects present in the labels.

Since LDA is an unsupervised machine learning method based only on the word frequencies in a document, the topic model cannot differentiate supplement indications from adverse events of supplement as it does not take context into account. For example, in Topic 12, *Cascara sagrada* was used as an anthranoid laxative to treat constipation [21], and celandine was indicated for spastic discomfort of the gastrointestinal tract [22]. The topic keywords, obstruction, intestinal obstruction, abdominal pain were not adverse events, but indications of these herbal supplements.

Our LDA analysis only considered the supplements with only one active ingredient entry. According to DSLD, more than 70% of the supplements contain at least two active ingredients. Therefore, it is of our interest to analyze this portion of the dietary supplements while addressing the entity recognition issue (the indications, the adverse events, and the interactions between ingredients) in the future. We may consider using existing knowledge base content to find the known interactions of ingredients in products.

Conclusion

In summary, we extracted the precautions statements from the labeling information of 2,014 dietary supplements that have only one active ingredient in DSLD. The MedDRA terms list was used to convert the statements into documents represented by a list of adverse events. The collection of supplement documents were analyzed by LDA topic models and eight out of the 50 resulting topics were evaluated by two human annotators. The product level topic relatedness ranged from 58.8% to 100%, while the ingredient level topic relatedness ranged from 57.1% to 100%. Five topics have shown considerable coherence among the topic keywords, and the representative ingredients have been demonstrated to be closely related to the topics. These results indicated that LDA could effectively group the dietary supplements by their adverse effects, and the grouping information could provide insight for further literature-based studies.

Acknowledgments

Research reported in this publication was supported by the National Center for Complementary & Integrative Health Award (R01AT009457) (Zhang) and the University of Minnesota Grant-In-Aid award (Zhang).

References

1. Kantor ED, Rehm CD, Du M, White E, Giovannucci EL. Trends in Dietary Supplement Use among US Adults from 1999–2012. JAMA. 2016; 316:1464–1474. [PubMed: 27727382]
2. Cohen PA. Assessing Supplement Safety – The FDA's Controversial Proposal. N Engl J Med. 2012; 366:389–391. [PubMed: 22276780]
3. Geller AI, Shehab ND, Weidle NJ, Lovegrove MC, et al. Emergency Department Visits for Adverse Events Related to Dietary Supplements. N Engl J Med. 2015; 373:1531–1540. [PubMed: 26465986]
4. Mason MJ. Drugs or Dietary Supplements: FDA's Enforcement of DSHEA. J Public Policy Mark. 1998; 17:296–302.
5. Frankos VH, Street DA, O'Neill RK. FDA Regulation of Dietary Supplements and Requirements Regarding Adverse Event Reporting. Clin Pharmacol Ther. 2009; 87:239–244. [PubMed: 20032973]
6. Denham BE. Dietary Supplements – Regulatory Issues and Implications for Public Health. JAMA. 2011; 306:428–429. [PubMed: 21730229]
7. Pittler MH, Schmidt K, Ernst E. Adverse events of herbal food supplements for body weight reduction: systematic review. Obes Rev. 2005; 6:93–111. [PubMed: 15836459]
8. Izzo AA, Ernst E. Interactions between herbal medicines and prescribed drugs: an updated systematic review. Drugs. 2009; 69:1777–1798. [PubMed: 19719333]
9. FDA regulations on supplement labeling. Available from: http://www.fda.gov/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/DietarySupplements/default.htm
10. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current aplications in bioinformatics. SpringerPlus. 2016; 5:1608. [PubMed: 27652181]
11. Blei DM. Probablistic Topic Models. Commun ACM. 2012; 55:77–84.
12. Sullivan R, Sarker A, O'Connor K, Goodin A, Karlsrud M, Gonzalez G. Finding potentially unsafe nutritional supplements from user reviews with topic modeling. Pac Symp Biocomput. 2016; 21:528–39. [PubMed: 26776215]
13. Bisgin H, Liu Z, Fang H, Xu X, Tong W. Mining FDA drug labels using an unsupervised learning technique – topic modeling. BMC Bioinformatics. 2011; 12:S11.
14. Bisgin H, Liu Z, Kelly R, Fang H, Xu X, Tong W. Investigating drug repositioning opportunities in FDA drug labels through topic modeling. BMC Bioinformatics. 2012; 13:S6.
15. Dietary Supplement Label Database. Available from: https://dsld.nlm.nih.gov/dsld
16. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016; 44:D1075–D1079. [PubMed: 26481350]
17. Rehurek, R., Sojka, P. Software Framework for Topic Modelling with Large Corpora. Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks; Valletta, Malta. 2010.
18. Pellitteri PK. Evaluation of hypercalcemia in relation to hyperparathyroidism. Otolaryngol Clin North Am. 2010; 43:389–397. [PubMed: 20510722]
19. Wolf ME, Moffat M, Mosnaim J, Dempsey S. Lithium therapy, hypercalcemia and hyperparathyroidism. Am J Ther. 1997; 4:323–325. [PubMed: 10423626]
20. DeLuca HF. The metabolism and functions of vitamin D. Adv Exp Med Biol. 1986; 196:361–75. [PubMed: 3012979]
21. Nakasone ES, Tokeshi J. A Serendipitous Find: A Case of Cholangiocarcinoma Identified Incidentally After Acute Liver Injury Due to *Cascara sagrada* Ingestion. Hawaii J Med Public Health. 2015; 74:200–202. [PubMed: 26114074]
22. Teschke R, Frenzel C, Glass X, Schulze J, Eickhoff A. Greater Celandine hepatotoxicity: a clinical review. Ann Hepatol. 2012; 11:838–48. [PubMed: 23109446]

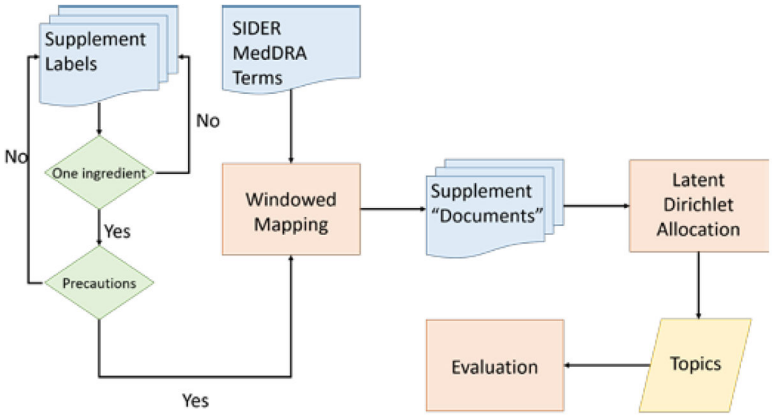

Figure 1.
Method overview flow chart

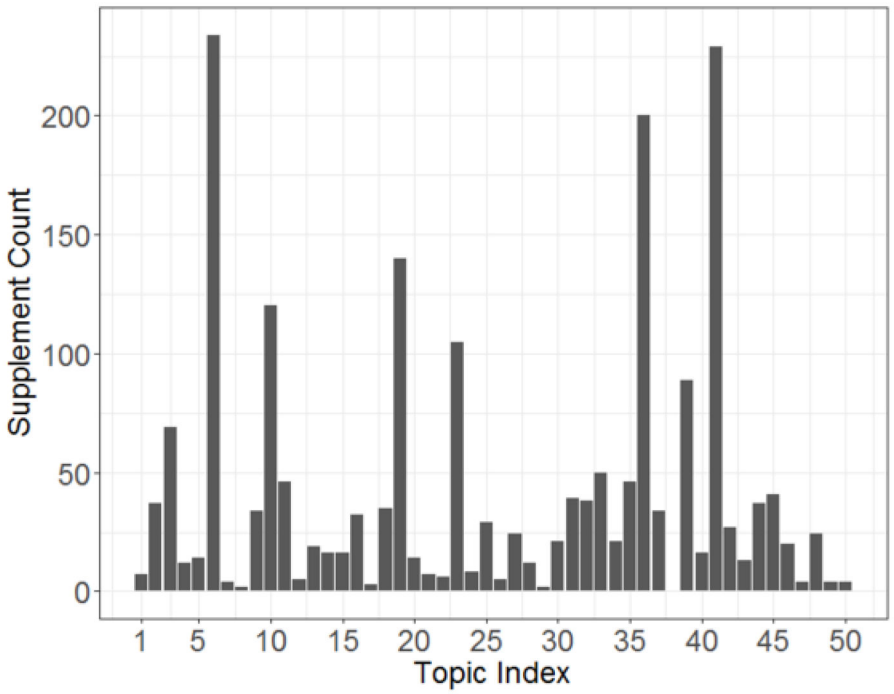

Figure 2.
The distribution of the supplement counts for the 50 topics

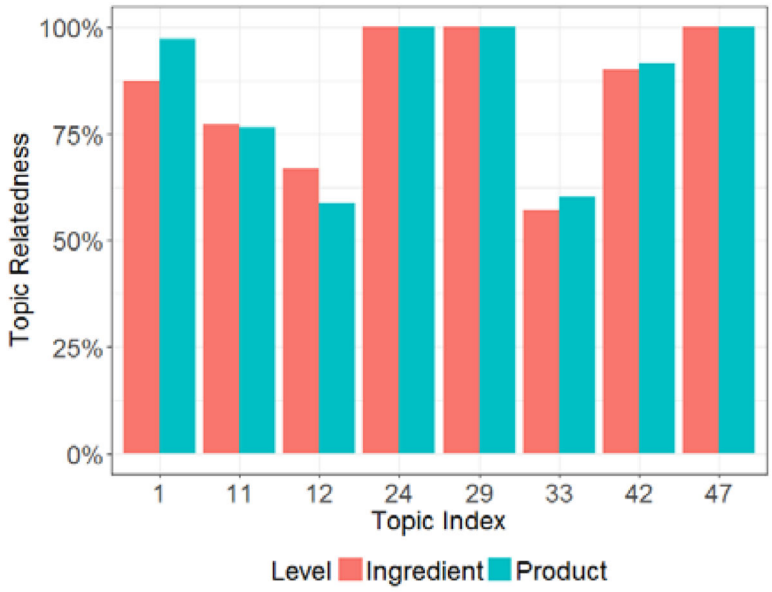

Figure 3.

Topic relatedness at both the product and the ingredient level of the eight topics listed in Table 1. The x-axis is labeled with the topic index.

Table 1

Topic keywords and the representative ingredients of eight topics chosen for manual evaluation. The topic keywords were listed in decreasing order by their probabilities generated by LDA. The bolded topics are coherent ones where the topic keywords are closely related.

Topic	Topic Keywords	Representative Ingredients
1	**Sarcoidosis, hyperparathyroidism, lymphoma, bipolar disorder**	Vitamin D3
11	**Anxiety, depression, anxiety depression, anxiety disorder**	S-Adenosyl-L-Methionine (SAMe), steroids
12	**Obstruction, inflammation, intestinal obstruction, pain, abdominal pain**	*Cascara sagrada*, Celandine
24	Hyperthyroidism, palpitations, duodenal ulcer, anxiety, blood pressure abnormal	L-Tyrosine, Iodine
29	**Gastritis, loose stools, peptic ulcer, heartburn, ulcer**	Oregano oil, Cayenne pepper
33	Photosensitivity, tingling skin, tingling sensation, bladder dysfunction, serotonin syndrome	Beta-Alanine, St. John's Wort
42	Epilepsy, blood disorder, cardiovascular disorder, hypotension, tuberculosis	St. John's Wort, Graviola
47	**Dizziness, headache, nervousness, glaucoma, depression**	Caffeine, L-Glutamine

Table 2

Comparison of the number of one-ingredient products and the number of unique ingredients in the topics under evaluation.

Topic	One-ingredient products	Unique Ingredients
1	37	8
11	44	34
12	17	6
24	26	7
29	18	6
33	20	7
42	12	10
47	19	12