Word-of-Mouth Innovation: Hypothesis Generation for Supplement Repurposing based on Consumer Reviews

Jung-wei Fan, PhD[1,2], Yves A. Lussier, MD[1,2,3]
[1]Department of Medicine; [2]Center for Biomedical Informatics & Biostatistics; [3]BIO5 Institute, The University of Arizona, Tucson, Arizona, USA

Abstract

Dietary supplements remain a relatively underexplored source for drug repurposing. A systematic approach to soliciting responses from a large consumer population is desirable to speed up innovation. We tested a workflow that mines unexpected benefits of dietary supplements from massive consumer reviews. A (non-exhaustive) list of regular expressions was used to screen over 2 million reviews on health and personal care products. The matched reviews were manually analyzed, and one supplement-disease pair was linked to biological databases for enriching the hypothesized association. The regular expressions found 169 candidate reviews, of which 45.6% described unexpected benefits of certain dietary supplements. The manual analysis showed some of the supplement-disease associations to be novel or in agreement with evidence published later in the literature. The hypothesis enrichment was able to identify meaningful function similarity between the supplement and the disease. The results demonstrated value of the workflow in identifying candidates for supplement repurposing.

Introduction

For more than a decade, there has been fervent interest in drug repurposing (or repositioning) due to the expensive and time-consuming process of drug development. It was estimated that on average it takes 10 years and at least $1 billion to bring a drug to market[1]. On the other hand, finding a new use for an old drug has advantages such as shortened time in verifying toxicity and higher probability of government approval. Renowned examples include sildenafil for erectile dysfunction and thalidomide for erythema nodosum leprosum. Interestingly, a relatively overlooked territory along this repurposing movement is dietary supplements, which are widely available off the shelf and may offer great variety of functions.

According to a 2015 survey by the Council for Responsible Nutrition (CRN)[2], about 68% of Americans use dietary supplements. Among these supplement users, 69% said their doctor talked to them about the benefits of taking supplements. Additionally, most supplement users aged 18-34 (66%) anticipate an increased use over the next 5 years. Dietary supplements come in diverse categories: minerals, vitamins, amino acids, enzymes, herbals and botanicals. These facts imply an extensive influence of dietary supplements over population health and well justify them as a rich source for discovering novel repurposing candidates. In fact, there has been ongoing "secondary repurposing" effort that searches for analogues of already repositioned chemicals (e.g., sildenafil[3,4]) in dietary supplements. There are tremendous opportunities for informatics to systematically facilitate the hypothesis generation process.

In our previous work of mining health-related issues in 1.3 million Amazon.com consumer reviews[5], we found that about 40% of the reviews described certain health benefits from using a grocery food product. The Amazon online store is one of the largest retailers in the United States, with over 300 million users and net sales of $136 billion (in 2016)[6]. As of 2017 March 6, its numbers of supplements sold by categories were: 19,619 Multi & Prenatal Vitamins, 16,048 (specific) Vitamins, 10,255 Minerals, 35,584 (non-herbal) Supplements, 42,519 Herbal Supplements, and 12,370 Weight Loss products. Given the huge collection of supplements and user base, the consumer feedback may actually serve as an ideal data source for automated surveillance. Currency is another merit of the data, as the reviews are constantly growing and closely reflect the products that people use.

Although there are issues with using consumer reviews (e.g., credibility and sparseness), we believe it is worth developing solutions that can bring any meaningful signal to our attention for advancing science. In this study, we proposed and exercised a dry run of a discovery workflow. Text mining was applied to screen for unexpected benefits of dietary supplements in the consumer reviews, followed by manual curation of the candidate reviews, and hypothesis enrichment by linking to external biological databases. The goal is two-fold: 1) assess content of the data for the target use in repurposing; and 2) execute the prototype workflow and identify issues for improvement.

Methods

We obtained a subset of 2,982,326 Amazon reviews on health and personal care products. The reviews were made available courtesy of McAuley et al.[7,8], who batch-fetched in previous work and shared the dataset with the research community. Our methods consisted of three stages, as elaborated below:

1. Text mining of reviews that mention an unexpected benefit: A java program was implemented to extract fields of interest (e.g., product ID, review text, review date) from the JSON file with ~2 million reviews and matched each review text with the following (case-insensitive) regular expression pattern:

   `\\b((unexpected|unintended|unanticipated|surprising)\\s+(effects?|benefits?))\\b`

2. Manual curation/analysis of the mined reviews: The regex-matched reviews were manually curated by the first author (JF) to determine which ones indeed described a certain unexpected benefit of a dietary supplement (or called the "true positive" reviews). The true positives were further analyzed for characterizing the types of the supplements and the benefits. The ones that did not meet the true positive criterion were also examined and categorized.

3. Hypothesis enrichment for specific findings of interest: To test the feasibility of generating richer hypotheses by knowledge integration, we selected a true positive review where the supplement and unexpected benefit (syndrome relief) were both unambiguous and searched them in biological databases. For the syndrome, we manually identified the corresponding trait in the Genome-Wide Association Studies (GWAS) Catalog[9] and its associated genes. As symptoms are also covered in the GWAS traits, this method is not limited to only diseases. For the supplement, the Comparative Toxicogenomics Database (CTD)[10] was manually looked up to find its associated genes. In addition, functional similarity of the genes between the supplement and the syndrome was then assessed by using an information-theoretic measure[11] to pinpoint possible biological explanation.

Results

Manual curation of the candidate reviews

Out of the 2,982,326 reviews, the regex pattern matched 169 candidate reviews. By reading through the candidate reviews, four categories were induced and summarized in **Table 1**. Less than half of the reviews (45.6%) contained unexpected benefits of interest (True positive) – with analysis elaborated in the next subsection. Among the true positives, there were five substance-effect associations supported by more than one review. For example, unexpected weight loss from using products with vitamin B2. A comparable portion (45.0%) of the reviews was about "non-dietary" health and personal care products (e.g., heart rate monitor and electric toothbrush), which were not filtered out in implementation of the screening program. The neutral effect category includes reviews with health-irrelevant property (e.g., enhanced flavor) or of debatable benefit (e.g., loss of appetite). The bottom minor category (Unspecified effect) is reviews with non-specific description such as "if you do a little research you may find some surprising benefits to xylitol". Although infrequent, those irrelevant cases indicate the regex-based approach may still pick up some noise.

Table 1. Categories identified from manually annotating the regex-matched candidate reviews

Category	Definition	Counts (%)
True positive	Dietary supplement with unexpected benefit indicated in the review	77 (45.6%)
Non-dietary	Product is health-related by not for dietary use	76 (45.0%)
Neutral effect	The effect from the dietary supplement is not apparently beneficial	10 (6.0%)
Unspecified effect	Description of specific effect is missing or unclear	6 (3.4%)

Qualitative analysis of the unexpected benefits

By inspecting the true positive reviews and product information, we found that about a half of the consumer-claimed "unexpected" benefits might actually be expected. Given the subjective nature, it is not uncommon that the reviewers were surprised due to lack of domain knowledge or just did not pay attention to the product labels. Nonetheless, there remained a good amount of informative reviews, from which we listed some notable repurposing contexts in **Table 2**. The examples demonstrate diversity of the potential uses, ranging from relief of symptoms, mental function enhancement, modification of substance use behavior, to deterring of infection vectors.

Table 2. Examples of potential supplement repurposing based on the reviews

Product ID	**Product type**	**Review excerpt**	**Review year**
B00008CQTS	Recombinant human growth hormone (hGH)	…very noticible increase in memory and cognitive skills	2006
B0000537A7	Hair regrowth tablets	…everyone around me was being bit and I wasn't bit once. The room I stayed in at night had a lot of mosquitos too. I killed over 20 of them in the room but I never suffered even one bite!	2011
B001DYKCJQ	Creatine monohydrate (for athlete performance enhancement)	…amazed how it controlled the acid reflux for me.	2012
B0058GXIYG	Ashwagandha extract capsules (marketed as a general healthy herb supplement)	I experienced an unexpected benefit by taking away the negative symptoms of IBS (with constipation) of which I have suffered for 33 years	2012
B0013OUKPC	Inositol powder (marketed for liver function support)	…a significant reduction in the number of heart palpitations (premature atrial contractions) I have.	2013
B0029O0RUS	Resveratrol (marketed as general healthy juice capsule)	A rather stubborn bout of eczema on my feet cleared up after being there for two years.	2013
B0087Q8GZA	Probiotics tablets (marketed as supplement for gut health)	…an unexpected benefit was the improvement in my sleep.	2013
B00EIW6NZC	L-arginine alpha-ketoglutarate (amino acid supplement)	A surprising effect is I quit smoking due to losing my desire for cigarettes since I started arginine.	2014

Hypothesis enrichment by linking to biological databases

From the true positive examples in **Table 2**, IBS-ashwagandha association (4th row under the heading) was utilized for the enrichment exercise, as the supplement and relief of the condition (IBS, irritable bowel syndrome) are both unambiguous according to the review. Ashwagandha (*Withania somnifera*) is a perennial shrub that has been used as medicinal herb in some cultures. To our knowledge, there is not any literature on the use of ashwagandha for IBS. We were able to identify their associated genes from the GWAS Catalog and CTD respectively (see **Table 3**). By further looking into the GO annotations[12], moderate functional similarity between the genes suggests possible biological underpinning of the observed effect. For example, **Figure 1** illustrates a couple of the GO Biological Process terms between SOD1 (ashwagandha) and PER2 (IBS), with information-theoretic similarity scores computed using our previously published method[11]. For GO Molecular Function terms, we also observed overlaps such as protein binding (GO:0005515) shared between PER2 (IBS) and HSPA9 (ashwagandha).

Table 3. Genes identified as associated with IBS and ashwagandha in the databases

IBS		**ashwagandha**	
Gene ID: 166378	Symbol: SPATA5	Gene ID: 847	Symbol: CAT
Gene ID: 55502	Symbol: HES6	Gene ID: 2670	Symbol: GFAP
Gene ID: 8864	Symbol: PER2	Gene ID: 3306	Symbol: HSPA2
Gene ID: 65217	Symbol: PCDH15	Gene ID: 3313	Symbol: HSPA9
Gene ID: 11014	Symbol: KDELR2	Gene ID: 4684	Symbol: NCAM1
		Gene ID: 6647	Symbol: SOD1

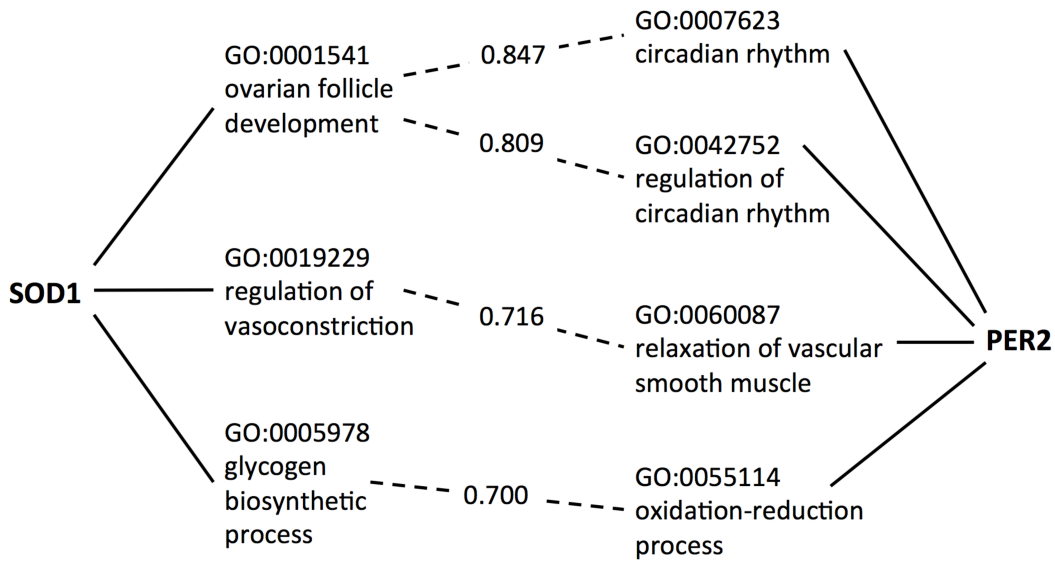

Figure 1. Similar biological processes between two genes associated with ashwagandha (SOD1) and IBS (PER2)

Discussion

Consumer reviews as innovation driver

Our results suggest that mining massive consumer reviews may enable (timely) discovery of useful incidents of unexpected dietary benefits. Enhanced with an integrative approach, rich hypotheses can be generated to spark novel research. Below we discuss several noteworthy cases from **Table 2**:

1. In the enriched analysis of **Figure 1**, the possible involvement of circadian rhythm and ovarian follicle development offers a biological hint that aligns with evidence of IBS gender disparity as reported in the

literature[13,14]. We believe systematically linking consumer experience to formal databases will shed light on potentially fruitful pharmacogenomics studies.

2. The product review on improvement in cognitive skills after taking recombinant human growth hormone was written in 2006, while the earliest literature we found regarding such association was published in 2013[15,16]. It is likely that screening from a large consumer population can help detect useful signals a few years ahead of other formal studies.
3. In 2013, a consumer reported eczema was resolved by dietary resveratrol, which does not seem to have been reported in literature. Interestingly, a US patent was filed in 2001 for external use of resveratrol for treating exfoliative eczema, acne, and psoriasis[17]. Due to less stringent regulation, this example suggests that supplement-based intellectual property could be nimbly registered as soon as a promising effect is found.
4. In 2013, a consumer reported unexpected sleep improvement by using probiotics tablets. Until 2014 and 2016 respectively, such beneficial effect on mice and human began surfacing in the literature[18,19]. Since many of the supplements are food-based, we foresee that mining the relevant consumer feedback would expedite hypothesis generation for the cutting-edge frontiers of microbiomics and broadly nutrigenomics.

Potential model of citizen science

The concept of citizen science[20,21] has been promoted in various domains. The basic idea is letting motivated lay people participate in research studies and perform tasks such as collecting samples or data. We believe that supplement repurposing is an ideal area to engage the general public into a crowd-sourced style of "observational trial". The advantages of leveraging massive review data include: diverse product types, large consumer base, and constant monitoring as integrated into daily life. To facilitate this mission, informatics should be able to contribute on vocabulary harmonization and streamlining the integration of data sources. An interesting observation was how some consumers described in reviews that they tested the effect via self-controlled experiment, i.e., verifying the effect by stopping use and resuming again. This suggests that, with moderate scientific training, we could further improve the quality of data collected from general consumers.

Limitations

As a proof-of-concept study, the review curation was conducted by only one annotator (JF) and without referring to a tested guideline. Due to the novelty in many of the findings, we could not find a proper reference standard for the evaluation. Some product reviews can be anecdotal/subjective by nature and of varying quality dependent on the consumer's background. It is not uncommon to find inconsistent comments on the supplement-effect from other Internet sources (e.g., the relation between creatine monohydrate and acid reflux). Those inconsistencies could be accounted by justifiable genetic variation or could be just noise from uncontrolled covariates. Multi-ingredient supplements also pose challenges in pinpointing the exact active element that caused the effect. A relevant argument is that the sense of "repurposing" actually becomes indistinct for supplements that are marketed as "all-purpose". In comparison to approved drugs, supplements also tend to conceal risks that are not well studied.

Future work

With observing a considerable amount of unwanted non-dietary product reviews, we will try to include a filter based on the finer category labels available in the detailed product information. A more rigorous annotation process will be needed, especially with a formal guideline and execution by multiple annotators. For higher sensitivity in detecting the semantics of unexpected benefits, we will explore natural language processing (NLP) and machine learning techniques. To automate the discovery workflow, a pipeline will be developed by incorporating the NLP and function similarity programs based on our previous work. The throughput can be boosted by implementing batch-lookup with the NCBI E-utilities[22] and directly processing downloaded GWAS Catalog and the CTD database. These enhancements should jointly contribute to a more accurate estimation of the yield (in terms of generating useful hypotheses) of the overall solution. Lastly, we still need to figure out a suitable reference standard for improving the validity and scalability of the evaluation.

Conclusion

To explore useful information for repurposing dietary supplements, we proposed a workflow of mining unexpected benefits mentioned in massive consumer reviews. The proof-of-concept study involved: 1) automatic matching of clue phrases to identify candidate reviews, 2) manual categorization/analysis of the candidate reviews, and 3) manually linking a novel supplement-disease pair to relevant biological databases for hypothesis enrichment. We found 45.6% of the 169 matched candidate reviews contained user-claimed unexpected benefits of dietary supplements, and about half of them were likely novel. By browsing relevant literature, some of the cases did demonstrate potential value for driving repurposing innovation. The hypothesis enrichment also derived informative functional associations between the supplement and the disease. More rigorous evaluation and validation approaches will be needed. The results show meaningful content in consumer reviews as well as the feasibility of a workflow to facilitate supplement repurposing.

Acknowledgement

We thank Colleen Kenost for assistance in editing the manuscript.

References

1. Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. 2010;9(3):203-214. doi:10.1038/nrd3078.
2. 2015 CRN consumer survey on dietary supplements. Council for Responsible Nutrition. http://www.crnusa.org/CRNconsumersurvey/2015/. Published 2015. Accessed March 5, 2017.
3. Wollein U, Eisenreich W, Schramek N. Identification of novel sildenafil-analogues in an adulterated herbal food supplement. *J Pharm Biomed Anal*. 2011;56(4):705-712. doi:10.1016/j.jpba.2011.07.012.
4. Ge X, Li L, Koh HL, Low MY. Identification of a new sildenafil analogue in a health supplement. *J Pharm Biomed Anal*. 2011;56(3):491-496. doi:10.1016/j.jpba.2011.06.004.
5. Torii M, Tilak SS, Doan S, Zisook DS, Fan J. Mining health-related issues in consumer product reviews by ssing scalable text analytics. *Biomed Inf Insights*. 2016;8(Suppl 1):1-11. doi:10.4137/BII.S37791.TYPE.
6. Smith C. 120 amazing Amazon statistics and facts (February 2017). DMR. http://expandedramblings.com/index.php/amazon-statistics/. Published 2017. Accessed March 6, 2017.
7. McAuley J, Targett C, Shi Q, Hengel A Van Den. Image-based recommendations on styles and substitutes. In: *Proceeding of 38th ACM SIGIR*. 2015:1-11. doi:10.1145/2766462.2767755.
8. McAuley J, Pandey R, Leskovec J. Inferring networks of substitutable and complementary products. In: *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'15)*. 2015:12. doi:10.1145/2783258.2783381.
9. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 2016:gkw1133. doi:10.1093/nar/gkw1133.
10. Mattingly CJ, Colby GT, Forrest JN, Boyer JL. The Comparative Toxicogenomics Database (CTD). *Environ Health Perspect*. 2003;111(6):793. doi:10.1289/ehp.6028.
11. Tao Y, Sam L, Li J, Friedman C, Lussier YA. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*. 2007;23(13):i529-38. doi:10.1093/bioinformatics/btm195.
12. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 2015;43(Database issue):D1049-56. doi:10.1093/nar/gku1179.
13. Herman J, Pokkunuri V, Braham L, Pimentel M. Gender distribution in irritable bowel syndrome is proportional to the severity of constipation relative to diarrhea. *Gend Med*. 2010;7(3):240-246.
14. Chang L, Heitkemper MM. Gender differences in irritable bowel syndrome. *Gastroenterology*. 2002;123(5):1686-1701. doi:10.1053/gast.2002.36603.
15. Watson SE, Rogol AD. Recent updates on recombinant human growth hormone outcomes and adverse events. *Curr Opin Endocrinol Diabetes Obes*. 2013;20(1):39-43. doi:10.1097/MED.0b013e32835b7ea8.
16. Wolfgram PM, Carrel AL, Allen DB. Long-term effects of recombinant human growth hormone therapy in children with Prader-Willi syndrome. *Curr Opin Pediatr*. 2013;25(4):509-514.
17. Pelliccia MT, Giannella A, Giannella J. Use of resveratrol for the treatment of exfoliative eczema, acne and

psoriasis. 2001. US Patent 20010056071.
18. Miyazaki K, Itoh N, Yamamoto S, et al. Dietary heat-killed Lactobacillus brevis SBC8803 promotes voluntary wheel-running and affects sleep rhythms in mice. *Life Sci*. 2014;111(1):47-52. doi:10.1016/j.lfs.2014.07.009.
19. Nakakita Y, Tsuchimoto N, Takata Y, Nakamura T. Effect of dietary heat-killed Lactobacillus brevis SBC8803 (SBL88^{TM}) on sleep: A non-randomised, double blind, placebo-controlled, and crossover pilot study. *Benef Microbes*. 2016;7(4):501-509. doi:10.3920/BM2015.0118.
20. Costandi M. Citizen microbiome. *Nat Biotechnol*. 2013;31(2):90. doi:10.1038/nbt0213-90a.
21. Bonney R, Shirk JL, Phillips TB, et al. Citizen science: Next steps for citizen science. *Science*. 2014;343(6178):1436-1437. doi:10.1126/science.1251554.
22. Entrez programming utilities help. https://www.ncbi.nlm.nih.gov/books/NBK25501/. Accessed June 19, 2017.