2015

A Computational Approach for Identifying Plant-Based Foods for Addressing Vitamin Deficiency Diseases

Christina Yu
University of Vermont

Indra Neil Sarkar
University of Vermont

A Computational Approach for Identifying Plant-Based Foods for Addressing Vitamin Deficiency Diseases

Christina Yu[1] and Indra Neil Sarkar[2,3]

Key words: computing methodologies, vitamin deficiency diseases, vegetarian diet

[1] Undergraduate Program in Biochemistry, University of Vermont, Burlington, VT
[2] Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT
[3] Center for Clinical and Translational Science, University of Vermont, Burlington, VT

ABSTRACT
Vitamins are nutrients that are essential to human health, and deficiencies have been shown to cause severe diseases. In this study, a computational approach was used to identify vitamin deficiency diseases and plant-based foods with vitamin content. Data from the United States Department of Agriculture Standard Reference (SR27), National Library of Medicine's Medical Subject Headings and MEDLINE, and Wikipedia were combined to identify vitamin deficiency diseases and vitamin content of plant-based foods. A total of 41,584 vitamin-disease associations were identified from MEDLINE-indexed articles as well as from entries in Wikipedia. The SR27 identified 1912 foods that contained at least one vitamin, with an average of 1276 foods per vitamin. Vitamin B12 and D contained the fewest number of foods (n=135 and 70, respectively). The results of this study establish the foundation for developing a process to link vitamin deficiency diseases to vitamin-rich foods.

INTRODUCTION

Vegetarian diets have been shown to have positive health benefits, as well as lower the overall risk of conditions like cardiovascular disease, type 2 diabetes, and cancer.[1] Changes in diet to more vegetable-based foods may have a direct impact on addressing conditions such as obesity, which has been noted as a risk factor for myriad conditions, such as cardiovascular disease and type 2 diabetes.[2] Such correlations support the notion that vegetable-based diets may be beneficial to maintaining better health, although some of the overall health benefit may be an artifact in that those who subscribe to a vegetarian diet may have a decreased consumption of foods that increase the risk of such conditions. Nonetheless, an analysis of the dietary patterns has shown that vegetarians consumed higher amounts of plant-based foods and lower amounts of added fats, sweets, snack foods, and non-water beverages compared to non-vegetarians.[3-5]

Vegetarian diets face the challenge of meeting daily nutritional needs that may be more readily fulfilled through non-vegetarian sources. Vitamins are organic compounds essential for the human body to function. They are obtained by ingesting certain foods, and a deficiency results when the body does not receive enough of a vitamin. Vitamin deficiencies can result in significant health problems if not properly supplemented. A study of adolescents in Sweden found that those who followed a vegan diet did not ingest enough of vitamin B12, which is not naturally present in plant-based foods.[6] Severe deficiency of vitamin B12 can result in neurological damage such as peripheral neuropathy, memory loss, and dementia.[7] Since vitamin B12 is not naturally found in plant-based foods, there can be higher risk of its deficiency for those following a strict vegetarian diet. Vitamin B12 deficiency may be addressed by consuming foods that have been fortified with vitamin B12 (e.g., certain cereals, soy and rice beverages, and nutritional yeast) or through daily supplementation.[8] In addition to addressing deficiencies of vitamins that may not be present in plant-based foods, it is also important to ensure that vegetarian diets have adequate vitamin content to support health and reduce the risk of conditions that may be due to vitamin-deficiency.

There are limited resources that catalogue vitamins in plant-based foods relative to vitamin deficiency conditions. The United States Department of Agriculture maintains a database of nutrient compositions of many foods available in the United States; the Standard Reference, Release 27 (SR27) contains information of up to 150 food components on 8,618 foods.[9] Using data from SR27, this study aimed to relate plant-based foods with vitamin content to common vitamin deficiency diseases based on scientific literature indexed in MEDLINE, the largest freely accessible citation database for biomedicine, and vitamin deficiency entries in Wikipedia, the largest general reference resource. The results of this study form the foundation for developing a resource to relate vitamin deficiency diseases with vitamin content of plant-based food, which may be used to support the development of nutrient balanced vegetarian diets.

Vitamin Background

Vitamin A

Vitamin A is fat-soluble, and the majority of the nutrient is stored in the liver. There are multiple isomers (vitamers) of vitamin A: retinol, retinal, retinoic acid, retinyl esters, and provitamin A carotenoids (i.e. precursors of retinol). Preformed vitamin A in the form of retinol is absorbed in the small intestine and the efficiency of absorption is between 70-90 percent. Carotenoids are also absorbed into the small intestine, but the efficiency is much lower, between 9-22 percent. The following carotenoids are converted to vitamin A in the body: α-carotene, β-carotene, and β-

cryptoxanthin. To reduce the risk of excess storage, vitamin A is excreted in the urine and bile. A consideration for vegetarians is that retinol is only found in animal-based foods. Therefore, vegetarians must consume provitamin A carotenoids to avoid deficiency. The recommended dietary allowances (RDA) for adult males and females (aged 19 and above) are 900 and 700 μg of retinol activity equivalents (RAE)/day, respectively.[10] Deficiencies in vitamin A may result in xerophthalmia, initially manifesting as night blindness and, in severe cases, resulting in keratomalacia.[11]

Vitamin B1

Also known as thiamin, vitamin B1 is a water-soluble nutrient that plays an important role in the metabolism of carbohydrates in the functional form of thiamin pyrophosphate.[7] Thiamin is mainly taken into the body through absorption in the small intestine. The RDA values for adult males and females are 1.2 and 1.1 mg/day, respectively.[10] The main sources of vitamin B1 include whole grains and pork, and the nutrient content can be significantly reduced with heating.[12] Deficiency results in symptoms of anorexia, weight loss, muscle weakness, decreased short-term memory, and cardiovascular effects. Diseases that are commonly seen from thiamin deficiency are beriberi and Wernicke-Korsakoff syndrome.[10]

Vitamin B2

Vitamin B2 is also known as riboflavin. It is a water-soluble nutrient that functions as a coenzyme in oxidation-reduction reactions in the predominant form of flavin adenine dinucleotide.[7] Absorption of the nutrient mainly occurs in the small intestine with a small amount occurring in the large intestine. The RDA values for adult males and females are 1.3 and 1.1 mg/day, respectively. Signs of vitamin B2 deficiency include glossitis, cheilosis, angular stomatitis, and are usually accompanied by other nutrient deficiencies.[10]

Vitamin B3

Commonly referred to as niacin, this nutrient is water-soluble and predominantly exists in bounds forms of nicotinamide and plays a large role in oxidation-reduction reactions during intracellular respiration and fatty acid synthesis.[7] Niacin is found in high concentrations in meat, whole grains, and fortified cereals. Some niacin can be synthesized by tryptophan with a relatively low efficiency of 60 μg tryptophan to 1 μg niacin. The RDA for adult males and females are 16 and 14 mg/day, respectively. Severe niacin deficiency results in pellagra and characterized by pigmented rash, depression, apathy, fatigue, headache, vomiting, constipation, or diarrhea.[10]

Vitamin B5

Also known as pantothenic acid, it is a water-soluble nutrient and a component of coenzyme A, which is involved in the synthesis of fatty acids and membrane phospholipids. Data were insufficient to calculate an RDA, thus adequate intake (AI) values were established instead. The AI for adult males and females is 5 mg/day. Sources of vitamin B5 include meats, whole grains, potatoes, egg yolks, and broccoli. Vitamin deficiency results in metabolic impairments and may include symptoms of irritability, restlessness, fatigue, apathy, malaise, nausea and vomiting, hypoglycemia, and other neurobiological symptoms.[10]

Vitamin B6

Vitamin B6 is water-soluble and consists of six vitamers: pyridoxal, pyridoxine, pyridoxamine, pyridoxal 5'-phosphate, pyridoxine 5'-phosphate, and pyridoxamine 5'-phosphate. Pyridoxal 5'-phosphate and pyridoxamine 5'-phosphate are the metabolically active forms of the vitamin and function as coenzymes in amino acid metabolism.[13] The RDA for adult males and females is 1.3 mg/day until age 30 and then increases to 1.7 and 1.5 mg/day for males and females, respectively. Sources of vitamin B6 include fortified cereals, meat, fish, poultry, starchy vegetables, and noncitrus fruits. Deficiency symptoms include seborrheic dermatitis, anemia, epileptiform convulsions, depression, and confusion.[10]

Vitamin B7

Also called biotin, nutritional data are insufficient for this nutrient. Biotin is water-soluble and known to function as a coenzyme in bicarbonate-dependent carboxylation reactions. In the small intestine, it is known to be bound to avidin, a protein in egg white, which then prevents the nutrient from being absorbed. The AI value for adult males and females is 30 μg/day. High concentrations of biotin are found liver. Deficiency signs include dermatitis, conjunctivitis, alopecia, depression, lethargy, and other central nervous system abnormalities.[10]

Vitamin B9

Known as folate, it functions as a coenzyme in nucleic and amino acid metabolism. The RDA for adult males and females is 400 μg/day, which can be supplied by fortified grain products, dark green vegetables, legumes, and other vegetables. The fortified form of vitamin B9 is referred to as folic acid. Alcohol, anticonvulsant drugs, cigarettes, and other drugs may impede folate activity. Deficiency symptoms include fatigue, irritability, headache, palpitations, and atrophic glossitis.[10]

Vitamin B12

Collectively known as cobalamins, the active forms of vitamin B12 are methylcobalamin and 5-deoxyadenosylcobalamin. This nutrient functions as a coenzyme that converts homocysteine to methionine, is required for red blood cell formation, and is involved in the metabolism of amino and fatty acids.[14] The RDA for men and women is 2.4 μg/day. Natural sources of vitamin B12 are found in animal products such as meats and dairy. However, many plant-based foods such as cereals have been fortified. Deficiency may result in hematological, neurological, and gastrointestinal effects.[10]

Vitamin C

Also called ascorbic acid, vitamin C is an antioxidant and functions as a cofactor in enzymatic and hormonal processes. Vitamin C is absorbed in the intestine via a sodium-dependent active transporter. The RDA values for males and females are 90 and 75 mg/day, respectively. Sources of vitamin C are citrus fruits, tomatoes, potatoes, and other fruits and vegetables.[10] The most common deficiency disease is scurvy; however, other symptoms include malaise, fatigue, and lethargy.[15]

Vitamin D

Collectively called calciferols, vitamin D can be obtained in two dietary forms: vitamin D2, known as ergocalciferol and derived from yeast and plants, and vitamin D3, known as cholecalciferol and found in animal products. Vitamin D can be synthesized in the skin through exposure to sunlight. In terms of dietary sources, the AI is 5 μg/day up through age 50 then increases to 10 μg/day for individuals through age 70 and 15 μg/day beyond age 70. Foods that naturally contain vitamin D include fatty fish and eggs from hens fed vitamin D. Most dietary intake comes from fortified dairy, cereal, and fruit juices. Deficiency results in osteoporosis, rickets in children, and osteomalacia in adults.[10]

Vitamin E

There are eight vitamers of vitamin E: α-, β-, γ-, and δ-tocopherols and α-, β-, γ-, and δ-tocotrienols. However, none of these vitamers are interconvertible, and only α-tocopherol is maintained in human plasma. The RDA for both males and females is 15 mg/day and is based on α-tocopherol intake. Vitamin E is mainly found in oils and fats, but can also be found in nuts, fruits, and vegetables. Deficiency signs include peripheral neuropathy, skeletal myopathy, and pigmented retinopathy.[10]

Vitamin K

The major form of dietary vitamin K is phylloquinone, which is derived from plants. Gut bacteria produce a form of vitamin K called menaquinone, but there is insufficient information on how this vitamer contributes to vitamin K status. The AI for males and females is 120 and 90 μg/day, respectively. Green vegetables and plant oils contain significant amounts of vitamin K. Deficiency may include hypoprothrombinemia, severe bleeding,[10] and radiographic knee osteoarthritis.[16]

MATERIALS AND METHODS

Vitamin Data

Data files were obtained from the United States Department of Agriculture (USDA) National Nutrient Database for Standard Reference, Release 27 (SR27; http://www.ars.usda.gov/Services/docs.htm?docid=8964). Plant-based foods were identified from SR27 as one of the following groups: "Spices and Herbs," "Fruits and Fruit Juices," "Vegetables and Vegetable Products," "Nut and Seed Products," "Legumes and Legume Products," and "Cereal Grains and Pasta." Vitamin-based nutritional information was also identified from within SR27 and consisted of all possible vitamer and vitamin entries.

Disease Data

Names of vitamins and corresponding vitamers were manually recorded from Wikipedia. These vitamin/vitamer names, along with the National Library of Medicine's Medical Subject Headings (MeSH) pertaining to disease (within the “C-Disease” hierarchy of MeSH), were used to query MEDLINE using the Entrez e-utilities to identify the number of disease articles associated with each vitamer or vitamin name. To calculate the importance of each vitamin-disease term, a weighted score called the term frequency-inverse document frequency (tf-idf) was utilized. The tf-idf is used to determine the importance of each vitamin-disease pair found in MEDLINE by proportionally weighting the number of times that the vitamin-disease term appears in the overall collection of research pertaining to the vitamin and all possible MeSH diseases. To calculate the tf-idf, the variables were defined as follows:

Term Frequency (TF)

The TF was defined as the number of articles containing each specific vitamin/vitamer-disease pair. For example, the TF for the vitamer "retinal" and the MeSH descriptor "Night Blindness" was queried as: *("retinal" NOT "vitamin A" NOT "retinol" NOT "retinoic acid" NOT "alpha-carotene" NOT "beta-carotene" NOT "gamma-carotene" NOT "beta-cryptoxanthin") AND "Night Blindness"[mh:exp]*. The term "retinal" was queried by itself without mention of the other vitamers of vitamin A (including the name of vitamin A) to identify articles that were solely about retinal in relation to night blindness. This way, comparisons could be made among different vitamers without the presence of duplicate articles. The MeSH component of the query used exploded MeSH hierarchies to catch all children terms. The above query found 429 articles on MEDLINE; therefore, the TF for "retinal" and "Night Blindess" is 429.

Document Frequency (DF)

The DF was defined as the number of articles that contained any mention of the vitamin/vitamer in relation to all possible MeSH diseases. Following the same vitamer-disease example from the TF definition, the respective DF query was *("retinal" NOT* <other vitamers of vitamin A>*) AND ("Arterial Occlusive Diseases"[mh:exp] OR "Arteriosclerosis"[mesh:exp] OR "Autoimmune Diseases"[mh:exp] OR…* <all possible MeSH>*)*. The term "retinal" was queried by itself to avoid duplicate articles when comparing amongst other vitamers, and all MeSH terms were exploded to include all children terms. This query found 93,170 articles on MEDLINE; therefore, the DF for "retinal" is 93,170.

Total Documents (N)

N was defined as the number of articles that contained mention of a vitamin/vitamer and a disease MeSH. The query consisted of all names of vitamins and vitamers connected by OR and all MeSH terms connected by OR. The vitamin and vitamer names were connected to the MeSH terms by AND. In query looked liked: *("vitamin A" OR* <all vitamers of vitamin A> *OR "vitamin B1" OR* <all vitamers of vitamin B1> *OR…* <all vitamins and vitamers>*) AND ("Arterial Occlusive Diseases"[mh:exp] OR "Arteriosclerosis"[mesh:exp] OR "Autoimmune Diseases"[mh:exp] OR…* <all possible MeSH>*)*. The number of articles returned was 270,475; therefore, N = 270,475.

The tf-idf was calculated for each vitamin/vitamer and MeSH disease using the following formula:

$$\text{tf-idf} = \log(\text{TF}) \times \log\left(\frac{\text{N}}{\text{DF}}\right)$$

Diseases pertaining to vitamin deficiencies were then manually identified from vitamin deficiency entries in Wikipedia. The disease names were mapped to MeSH descriptors using the National Center for Biomedical Ontology Annotator, where matches were identified as the longest term in MeSH. Each NCBO Annotator predicted mapping was manually verified. For each Wikipedia based vitamin-disease pair, the corresponding tf-idf score from the MEDLINE determined relationship was increased by one.

RESULTS

USDA Standard Reference

Nutrient values were reported in SR27 per 100 g of the food. Each nutrient consisted of at least one standard reporting unit. The SR27 did not include information on biotin for any plant-based foods.

Vitamin A

The SR27 reported vitamin A content in multiple forms - international units (IU) and μg of the following: RAE, retinol, α-carotene, β-carotene, and β-cryptoxanthin. Standard vitamin A activity was reported as the RAE value. IUs were also included since it is the commonly utilized in food labels. Conversion from IU to RAE used the equivalency that 1 IU is equivalent to 0.3 μg RAE. Retinol values were a straight 1:1 comparison to RAEs, and the carotenoid (α-carotene, β-carotene, and β-cryptoxanthin) values were converted using the following equivalencies provided in the SR27 documentation:[9]

$$1\ \mu g\ RAE = 24\ \mu g\ \alpha\text{-carotene} = 12\ \mu g\ \beta\text{-carotene} = 24\ \beta\text{-cryptoxanthin}$$

Vitamin B9

The standard reporting of vitamin B9 was in μg of dietary folate equivalents (DFEs). The SR27 reported μg of folate content in four forms: total folate, folic acid, food folate, and DFE folate. The naturally occurring form of the nutrient was referred to as food folate while the monoglutamate form used in fortified foods was referred to as folic acid. The sum of the values from food folate and folic acid resulted in the measure of total folate. To convert the different folate sources to DFE, the following conversion was provided in the SR27 documentation:[9]

$$1\ DFE = 1\ \mu g\ \text{food folate} = 0.6\ \mu g\ \text{folic acid}$$

Vitamin B12

Vitamin B12 was reported in units of μg as a naturally occurring nutrient and as a fortified nutrient. The naturally occurring form was the standard reporting of the vitamin. Since plants do not naturally contain vitamin B12, the fortified value was reported instead.

Vitamin D

The standard reporting of vitamin D was in IU. The SR27 further reported values of vitamins D, D2, and D3 in μg where the composition of vitamin D was defined as the sum of vitamin D2 and vitamin D3. Values of vitamin D were converted to IUs with the conversion factor detailed by the SR27 report that 1 μg of any form of vitamin D is equivalent to 40 IUs.[9]

Other SR27 considerations

Vitamin B9 conversions did not include "total folate" because the value is accounted for by conversion of food folate and folic acid. Certain vitamers were unable to be converted to comparable vitamin activity values. This describes the vitamin E isomers due to inability of the human liver to convert β-tocopherol, γ-tocopherol, δ-tocopherol, α-tocotrienol, β-tocotrienol, γ-tocotrienol, and δ-tocotrienol into the active α-tocopherol. Likewise, vitamin K isomers menaquinone and dihydrophylloquinone were unable to be made comparable to the activity of phylloquinone due to incomplete knowledge of their contribution to vitamin K status.

A total of 1912 foods were found to contain at least one vitamin. Table 1 summarizes the number of foods associated with each vitamin and vitamer; vitamin B2 contained the most number of foods (n = 1834) while vitamin D contained the least (n = 70). The average number of foods per vitamin was 1276 without accounting for vitamers. The addition of vitamer information did not contribute many new foods to the standard reporting results, with the exception of vitamin A, where 299 additional results were observed.

The overall composition of nutrients by food group observed was 2.90% Spices and Herbs, 19.2% Fruits and Fruit Juices, 47.3% Vegetables and Vegetable Products, 6.60% Nut and Seed Products, 15.2% Legumes and Legume Products, and 8.70% Cereal Grains and Pasta (Table 2). The majority of foods for each vitamin followed the overall food group composition. Outliers were vitamin B12 (n=135) and vitamin D (n=70). Both vitamins did not contain any foods with vitamin activity within the Spices and Nuts food groups, but contained more foods in the Legumes food group. Furthermore, vitamin B12 contained significantly less foods in the Fruits and Vegetable groups. Table 3 shows a sample of twenty foods containing the most vitamin B12. All of these foods are meat substitute products catalogued within the Legume and Legume Products food group. Data for all vitamins can be found in the supplementary data.

Table 1. Counts of plant-based foods containing each vitamin from SR27.

Vitamin	**Standard Reporting**	**With Vitamers**
A	1088	1387
B1	1808	-
B2	1834	-
B3	1822	-
B5	1637	-
B6	1796	-
B9	1649	1723
B12	135	135
C	1427	-
D	70	70
E	1104	1112
K	942	956

Vitamin B7 information was not available in the SR27.
Vitamins B1, B2, B3, B5, B6, and C did not report vitamer or fortification information.

Table 2. Counts of plant-based foods by vitamin content and food group from SR27.

	Food Group					
Vitamin	Spices and Herbs	Fruits and Fruit Juices	Vegetables and Vegetable Products	Nut and Seed Products	Legumes and Legume Products	Cereal Grains and Pasta
Vitamin A	44	264	594	68	81	37
Vitamin A with Vitamers	48	307	694	97	170	71
Vitamin B1	55	327	793	131	321	181
Vitamin B2	54	331	803	126	341	179
Vitamin B3	55	335	800	131	322	179
Vitamin B5	29	292	782	128	232	174
Vitamin B6	54	323	794	130	314	181
Vitamin B9	47	262	784	128	253	175
Vitamin B9 with Vitamers	47	278	798	131	293	176
Vitamin B12	0	4	16	0	102	13
Vitamin C	53	333	798	94	144	5
Vitamin D	0	2	32	0	31	5
Vitamin E	28	239	506	65	144	122
Vitamin E with Vitamers	30	239	511	66	144	122
Vitamin K	26	201	487	33	114	81
Vitamin K with Vitamers	26	201	494	33	114	88
Total	596	3938	9686	1361	3120	1789
Proportion of Food Group	0.029	0.192	0.473	0.066	0.152	0.087

Table 3. Top twenty foods with the most vitamin B12 content per 100 g from SR27.

Food	Food Group	Nutrient Value	Nutrient Units
LOMA LINDA Big Franks, canned, unprepared	Legumes and Legume Products	13.1	ug
MORNINGSTAR FARMS Sausage Style Recipe Crumbles, frozen, unprepared	Legumes and Legume Products	8.4	ug
MORNINGSTAR FARMS Grillers Burger Style Recipe Crumbles, frozen, unprepared	Legumes and Legume Products	8.3	ug
MORNINGSTAR FARMS Grillers Quarter Pound Veggie Burger, frozen, unprepared	Legumes and Legume Products	7.7	ug
MORNINGSTAR FARMS Breakfast Sausage Links, frozen, unprepared	Legumes and Legume Products	7.4	ug
LOMA LINDA Swiss Stake with Gravy, canned, unprepared	Legumes and Legume Products	7.3	ug
WORTHINGTON Prosage Links, frozen, unprepared	Legumes and Legume Products	6.7	ug
LOMA LINDA Vege-Burger, canned, unprepared	Legumes and Legume Products	6.6	ug
MORNINGSTAR FARMS Breakfast Sausage Patties Maple Flavored, frozen, unprepared	Legumes and Legume Products	6.6	ug
Meat extender	Legumes and Legume Products	6	ug
WORTHINGTON Smoked Turkey Roll, frozen, unprepared	Legumes and Legume Products	5.5	ug
MORNINGSTAR FARMS Hot and Spicy Veggie Sausage Patties, frozen, unprepared	Legumes and Legume Products	5.5	ug
Chicken, meatless, breaded, fried	Legumes and Legume Products	5.11	ug
MORNINGSTAR FARMS Grillers Original, frozen, unprepared	Legumes and Legume Products	4.5	ug
WORTHINGTON Vegetarian Burger, canned, unprepared	Legumes and Legume Products	4.4	ug
WORTHINGTON Vegetable Steaks, canned, unprepared	Legumes and Legume Products	4.2	ug
Fish sticks, meatless	Legumes and Legume Products	4.2	ug
Vegetarian fillets	Legumes and Legume Products	4.2	ug
Luncheon slices, meatless	Legumes and Legume Products	4	ug
WORTHINGTON Low Fat Veja-Links, canned, unprepared	Legumes and Legume Products	3.9	ug

Vitamin Disease Associations

The full data set of MEDLINE and Wikipedia queries consisted of 41,584 vitamin-disease associations with the combined scores ranging from 0 to 6.941 (Mean = 1.072, SD = 1.153). The MEDLINE data contains all 41,584 vitamin-disease associations and has the same range as the combined data, but with slightly lower statistics (Mean = 1.068, SD = 1.146). The Wikipedia data consists of 182 associations with scores that range from 0 to 6.918 with a higher mean (Mean = 3.163, SD = 1.506). The score quantifies the relative importance between the vitamin and disease term; therefore, a higher score means a greater vitamin-disease relationship. Figure 1 shows the kernel density estimation of the score from each data source. A kernel density estimation is a method of predicting the probability density function of a set of data. The area under the curve of a density estimation graph gives the relative likelihood of a certain vitamin-disease score occurring. Looking at the distribution of MEDLINE data, there is a large density peak at zero and most of the vitamin-disease scores fall between zero and two. The Wikipedia distribution shows a broader spread centered around three. The combined data from MEDLINE and Wikipedia shows a distribution that is very similar to the MEDLINE data. The top 20 vitamer-disease associations are displayed in Table 5. The full association data can be seen in supplementary data.

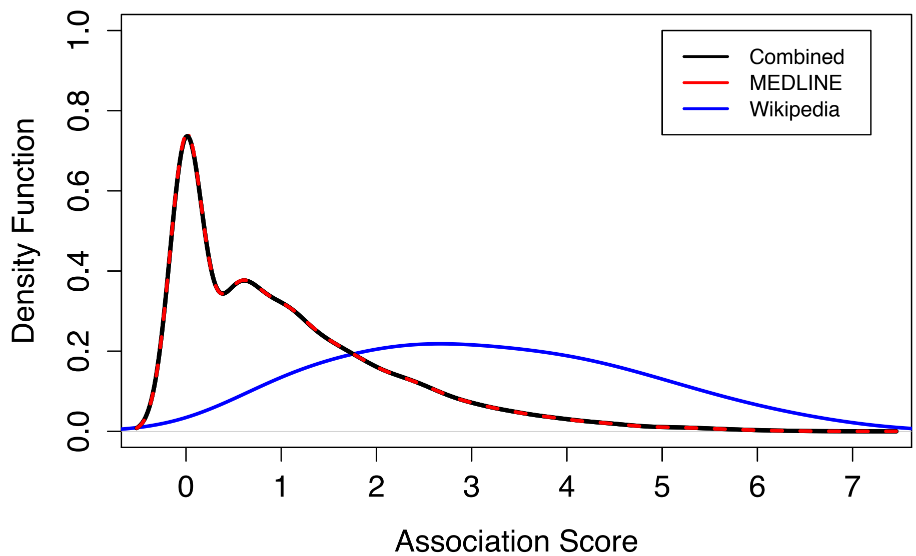

Figure 1. Kernel density plot of vitamin-disease association scores obtained by combining tf-idf scores from MEDLINE with a score of one for each vitamin-disease term match on Wikipedia. The area under the curve shows the relative likelihood of observing a certain combined association score.

Table 5. Twenty most relevant vitamin-disease associations in MEDLINE and Wikipedia.

Vitamer	Vitamin	Disease	TF	DF	N	Weighted score	Wikipedia score	Combined score
folinic acid	B9	Neoplasms	1489	1756	270475	6.941	0	6.941
methylcobalamin	B12	Nervous System Diseases	62	135	270475	5.918	1	6.918
folinic acid	B9	Neoplasms by Site	1335	1756	270475	6.837	0	6.837
cyanocobalamin	B12	Vitamin B 12 Deficiency	76	267	270475	5.653	1	6.653
folinic acid	B9	Digestive System Diseases	1097	1756	270475	6.651	0	6.651
folinic acid	B9	Digestive System Neoplasms	1075	1756	270475	6.632	0	6.632
folinic acid	B9	Gastrointestinal Diseases	975	1756	270475	6.539	0	6.539
folinic acid	B9	Gastrointestinal Neoplasms	958	1756	270475	6.522	0	6.522
riboflavin	B2	Riboflavin Deficiency	846	3557	270475	5.507	1	6.507
vitamin B6	B6	Nutritional and Metabolic Diseases	549	1193	270475	6.453	0	6.453
folinic acid	B9	Intestinal Diseases	810	1756	270475	6.363	0	6.363
folinic acid	B9	Intestinal Neoplasms	804	1756	270475	6.356	0	6.356
vitamin B12	B12	Vitamin B 12 Deficiency	3110	7963	270475	5.348	1	6.348
folinic acid	B9	Colorectal Neoplasms	797	1756	270475	6.347	0	6.347
cyanocobalamin	B12	Nervous System Diseases	60	267	270475	5.344	1	6.344
pyridoxamine	B6	Diabetes Mellitus	78	124	270475	6.317	0	6.317
pyridoxamine	B6	Endocrine System Diseases	78	124	270475	6.317	0	6.317
vitamin B6	B6	Nutrition Disorders	472	1193	270475	6.298	0	6.298
folinic acid	B9	Colonic Diseases	742	1756	270475	6.279	0	6.279
cyanocobalamin	B12	Nutritional and Metabolic Diseases	120	267	270475	6.249	0	6.249

TF = term frequency of the specified vitamer and disease pair

N = total number of documents that contain any vitamer and disease

DF = document frequency of the specified vitamer with any disease

Weighted score = tf-idf for the vitamer and disease pair; tf-idf = log(TF) × log(N/DF)

Wikipedia score = 1 if there was a match with vitamin and disease pair on Wikipedia; zero otherwise

Combined score = sum of weighted score and Wikipedia score

DISCUSSION

Eating a vegetarian diet can decrease the risk of health diseases, but the possibility of diseases caused by vitamin deficiency is increased. To identify plant-based foods for supplementing these deficiencies, data were relationally mapped from MEDLINE, the USDA SR27, and Wikipedia. The relationship data generated becomes the basis for an interactive search system that can be utilized by the general public and health professionals.

The findings of this study reveal that plant-based foods are good sources of all vitamins except for B12 and D. This result is not unexpected since vitamin B12 is naturally found in animal products and vitamin D is primarily obtained through exposure to sunlight.[10] Plant-based foods with these nutrients are often fortified to address these deficiencies.

The vitamin deficiency diseases identified by Wikipedia showed that the subset of vitamin-disease term scores had a higher mean than the full set of terms from MEDLINE (Mean = 3.163, 1.068, respectively). The probability density showed that a vitamin-disease term identified from Wikipedia was more likely to have a significant score and thus a stronger relationship. This means that Wikipedia was successful in finding significantly related terms from the large collection of MEDLINE-indexed literature.

The challenge of using Wikipedia as a resource is that it is not peer reviewed in the same way as the biomedical literature indexed in MEDLINE. As such, Wikipedia may contain terms or phrases that are not used in biomedical literature. For example, Wikipedia lists ε-tocopherol and ε-tocotrienol as isomers of vitamin E; however, there is no mention of these compounds in the SR27 or fact sheet for health professionals from the NIH.[17]. Since no reference was included for the ε-vitamers, the statement cannot be verified. An exact term search for ε-tocopherol and ε-tocotrienol on PubMed resulted in no matches. Further investigation resulted in discovering that ε-tocopherol was wrongly identified and is instead a derivative of β-tocopherol.[18]

Nonetheless, the distinguishing factor of Wikipedia is that it is an open encyclopedia that is readily accessible to and modified by the public. Therefore any information presented on the site will be seen by many and can be used to gauge public awareness of scientific research. To prevent controversial findings, Wikipedia strongly recommends the use of secondary sources due to the possibility of non-reproducible experimental results of primary sources.[19] For example, three review articles published within the past five years were cited in regards to folate deficiency under the section heading of the same name.[20] Since these references are fairly recent, it suggests that the topics of folate deficiency, and folate in general, are still relevant to the public. However, many of the disease symptoms associated with folate deficiency poorly referenced a pharmaceutical company; the reference linked to the company's main website, and some further probing on the company website was needed to find direct information on folate deficiency. Misinformation will be present on Wikipedia, but it can still be used as a research tool to identify diseases related to vitamin deficiency.

A significant contribution of this study is that it demonstrates the potential to connect disease and food data knowledge. Ultimately, these linkages may enable the identification of vitamins in plant-based foods that can address specific vitamin deficiency conditions. Tables 3 and 4 directly illustrate this association for vitamin B12. Cyanocobalamin and vitamin B12 were referenced in journal articles pertaining to vitamin B12 deficiency and further corroborated by information on Wikipedia. This deficiency can be supplemented by foods high in vitamin B12 such as meat substitutes from companies like Loma Linda, Morningstar Farms, and Worthington. This resource is especially useful for vegetarians and vegans as a quick method of determining if they are consuming an appropriate amount of nutrients.

Future directions will focus on the creation of an interactive system that uses disease and food queries to give vitamin-disease associations with specific regard to vitamin deficiencies and the plant-based foods that may supplement the deficiency.

FUNDING

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011963. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the National Institutes of Health.

REFERENCES

1. Position of the American Dietetic Association: Vegetarian Diets. *Journal of the American Dietetic Association*. 2009;109(7):1266-1282.
2. Gittelsohn J, Wolever TM, Harris SB, Harris-Giraldo R, Hanley AJ, Zinman B. Specific patterns of food consumption and preparation are associated with diabetes and obesity in a Native Canadian community. *The Journal of nutrition*. Mar 1998;128(3):541-547.
3. Orlich MJ, Jaceldo-Siegl K, Sabate J, Fan J, Singh PN, Fraser GE. Patterns of food consumption among vegetarians and non-vegetarians. *The British journal of nutrition*. Nov 28 2014;112(10):1644-1653.
4. Butler TL, Fraser GE, Beeson WL, et al. Cohort profile: The Adventist Health Study-2 (AHS-2). *International journal of epidemiology*. Apr 2008;37(2):260-265.
5. Rizzo NS, Jaceldo-Siegl K, Sabate J, Fraser GE. Nutrient profiles of vegetarian and nonvegetarian dietary patterns. *Journal of the Academy of Nutrition and Dietetics*. Dec 2013;113(12):1610-1619.
6. Larsson CL, Johansson GK. Dietary intake and nutritional status of young vegans and omnivores in Sweden. *The American journal of clinical nutrition*. 2002;76(1):7.
7. Combs GF, Jr. *The Vitamins: Fundamental Aspects in Nutrition and Health*. San Diego, California: Academic Press, Inc.; 1992.
8. Craig WJ. Health effects of vegan diets. *The American journal of clinical nutrition*. May 2009;89(5):1627S-1633S.
9. US Department of Agriculture ARS, Nutrient Data Laboratory. USDA National Nutrient Database for Standard Reference, Release 27. 2014.
10. Otten JJ, Hellwig JP, Meyers LD. *DRI, dietary reference intakes : the essential guide to nutrient requirements*. Washington, D.C.: National Academies Press; 2006.
11. Sommer A. Vitamin a deficiency and clinical disease: an historical overview. *The Journal of nutrition*. Oct 2008;138(10):1835-1839.
12. NIH. Thiamin. 2015. http://ods.od.nih.gov/factsheets/Thiamin-HealthProfessional/ Accessed April 2015.
13. NIH. Vitamin B6. 2011. http://ods.od.nih.gov/factsheets/VitaminB6-HealthProfessional/. Accessed April 2015.
14. NIH. Vitamin B12. 2011. http://ods.od.nih.gov/factsheets/VitaminB12-HealthProfessional/. Accessed April 2015.
15. Lykkesfeldt J, Michels AJ, Frei B. Vitamin C. *Advances in nutrition*. Jan 2014;5(1):16-18.
16. Misra D, Booth SL, Tolstykh I, et al. Vitamin K deficiency is associated with incident knee osteoarthritis. *The American journal of medicine*. Mar 2013;126(3):243-248.
17. NIH. Vitamin E. 2013. http://ods.od.nih.gov/factsheets/VitaminE-HealthProfessional/. Accessed April 2015.
18. Bunyan J, Mc HD, Green J, Marcinkiewicz S. Biological potencies of epsilon- and zeta-1-tocopherol and 5-methyltocol. *The British journal of nutrition*. 1961;15:253-257.
19. Wikipedia. Wikipedia:Identifying reliable sources (medicine). http://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources_(medicine). Accessed April 2015.
20. Wikipedia. Folic Acid. http://en.wikipedia.org/wiki/Folic_acid. Accessed September 2014.

21. Barnard ND, Bush AI, Ceccarelli A, et al. Dietary and lifestyle guidelines for the prevention of Alzheimer's disease. *Neurobiology of aging*. Sep 2014;35 Suppl 2:S74-78.
22. Robinson E, Harris E, Thomas J, Aveyard P, Higgs S. Reducing high calorie snack food in young adults: a role for social norms and health based messages. *Int J Behav Nutr Phys Act*. 2013;10:73.
23. Orlich MJ, Fraser GE. Vegetarian diets in the Adventist Health Study 2: a review of initial published findings. *The American journal of clinical nutrition*. Jul 2014;100 Suppl 1:353S-358S.
24. Piccoli GB, Clari R, Vigotti FN, et al. Vegan-vegetarian diets in pregnancy: danger or panacea? A systematic narrative review. *BJOG : an international journal of obstetrics and gynaecology*. Apr 2015;122(5):623-633.
25. Pawlak R, Lester SE, Babatunde T. The prevalence of cobalamin deficiency among vegetarians assessed by serum vitamin B12: a review of literature. *European journal of clinical nutrition*. May 2014;68(5):541-548.
26. Watanabe F, Yabuta Y, Bito T, Teng F. Vitamin B(1)(2)-containing plant food sources for vegetarians. *Nutrients*. May 2014;6(5):1861-1873.
27. Woo KS, Kwok TC, Celermajer DS. Vegan diet, subnormal vitamin B-12 status and cardiovascular health. *Nutrients*. Aug 2014;6(8):3259-3273.
28. Zhang R, Naughton DP. Vitamin D in health and disease: current perspectives. *Nutrition journal*. 2010;9:65.
29. Booth SL, Suttie JW. Dietary intake and adequacy of vitamin K. *The Journal of nutrition*. May 1998;128(5):785-788.
30. Suttie JW, Booth SL. Vitamin K. *Advances in nutrition*. Sep 2011;2(5):440-441.
31. Wikipedia. Vitamin A. http://en.wikipedia.org/wiki/Vitamin_A. Accessed September 2014.
32. Wikipedia. Thiamine. http://en.wikipedia.org/wiki/Thiamine. Accessed September 2014.
33. Wikipedia. Riboflavin. http://en.wikipedia.org/wiki/Riboflavin. Accessed September 2014.
34. Wikipedia. Niacin. http://en.wikipedia.org/wiki/Niacin. Accessed September 2014.
35. Wikipedia. Pantothenic Acid. http://en.wikipedia.org/wiki/Pantothenic_acid. Accessed September 2014.
36. Wikipedia. Vitamin B6. http://en.wikipedia.org/wiki/Vitamin_B6. Accessed September 2014.
37. Wikipedia. Biotin. http://en.wikipedia.org/wiki/Biotin. Accessed September 2014.
38. Wikipedia. Vitamin B12. http://en.wikipedia.org/wiki/Vitamin_B12. Accessed September 2014.
39. Wikipedia. Vitamin C. http://en.wikipedia.org/wiki/Vitamin_C. Accessed September 2014.
40. Wikipedia. Vitamin D. http://en.wikipedia.org/wiki/Vitamin_D. Accessed September 2014.
41. Wikipedia. Vitamin E. http://en.wikipedia.org/wiki/Vitamin_E. Accessed September 2014.
42. Wikipedia. Vitamin K. http://en.wikipedia.org/wiki/Vitamin_K. Accessed September 2014.
43. Wikipedia. Vitamin. http://en.wikipedia.org/wiki/Vitamin. Accessed September 2014.