Identifying Plant-Human Disease Associations in Biomedical Literature: A Case Study

Vivekanand Sharma, PhD[1], Wayne Law, PhD[2], Michael J. Balick, PhD[2], Indra Neil Sarkar, PhD, MLIS[1]
[1]Center for Biomedical Informatics, Brown University, Providence, RI USA
[2]Institute of Economic Botany, The New York Botanical Garden, Bronx, NY USA

Abstract

The impact of ethnobotanical data from surveys of traditional medicinal uses of plants can be enhanced through the validation of biomedical knowledge that may be embedded in literature. This study aimed to explore the use of informatics approaches, including natural language processing and terminology resources, for extracting and comparing ethnobotanical leads from biomedical literature indexed in MEDLINE. Using ethnobotanical data for plant species described in Primary Health Care Manuals of the Micronesian islands of Palau and Pohnpei, the results of this study were done relative to disease concepts from the "Mental, Behavioral And Neurodevelopmental Disorders" ICD-9-CM category. The results from this feasibility study suggest that informatics methods can be used to extract and prioritize relevant ethnobotanical information from biomedical knowledge literature.

Introduction

Understanding the use of plants by cultures around the world, the science of ethnobotany, has been shown as a potential source for the identification of new therapies (1, 2). Depending on the disease state being studied, searches for bioactive components that are driven by ethnobotanical knowledge are more effective than random selection of plant species (3); however, the process of cataloguing putative medicinal uses of plants, such as those that are compiled in field surveys, often depends on a labor intensive analysis of previous related ethnobotanical explorations and biomedical literature (4). Extensive searches of biomedical literature are also an essential aspect of the ethnobotanical pipeline for documenting and evaluating the pharmacological relevance of collected indigenous information. The extraction of meaningful information from ethnobotanical and biomedical texts thus remains an essential task in the cataloguing of plants with potential medicinal properties (5).

Ethnobotanical research focuses on organizing information, including therapeutic applications, about the historical and contemporary interactions between plants and traditional societies (6). Ethnobotanists aim to capture traditional botanical knowledge by interacting with people who are knowledgeable about plant uses, often but not limited to indigenous populations. Various qualitative and quantitative methods are employed to gather useful evidence pertaining to plant use patterns. Plant voucher specimens are also collected and taxonomic studies conducted for correct identification and nomenclature of plant species (7). Subsequent validation of potential plants with therapeutic properties is accomplished through a combination of systematic searches of literature, existing collections of herbaria (collections of preserved plant specimens), and ethnobotanical surveys (8). The study of the diversity of plants and their uses across various cultures may reveal essential patterns that provide insights to potential therapies. For example, a cross-cultural ethnomedicinal evaluation revealed similar uses of phylogenetically related plant species (9). However, to date, the adoption of informatics approaches have been slow in their application in the field of ethnobotany (10).

There have been significant advances in Natural Language Processing (NLP) tools and techniques for addressing the issues of variability, ambiguity, and context-dependent interpretation (11). Tools like MetaMap (12) and the National Center for Biomedical Ontology (NCBO) Annotator (13) have been shown to be effective in identifying biomedical concepts from free text (14, 15), including the extraction and ranking of key associations between biomedical entities from biomedical corpora (16-19). Several approaches have been developed for automating the extraction of entity relations and inferring new or hidden relations from biomedical text. Co-occurrence statistics based methods are commonly used to extract relations among biomedical entities (20). This approach relies on the assumption that two biomedical entities co-occur within the scope of a given text are likely to be related (21). In addition to the co-occurrence based methods, rule-based, statistical, machine learning and NLP based methods have also been used to extract relations from biomedical literature (22-26). For example, systems like BioMedLEE (27) and SemRep (28) have been developed for extracting relations between entities using syntactic and linguistic

analysis. Use of such systems for integrating semantic relations with co-occurrence have also been explored (29). Methods commonly used for extracting entity relations from text has been reviewed by Zweigenbaum, *et al.* (30).

This preliminary study aimed to demonstrate the potential to leverage existing NLP approaches for extracting and ranking disease associations for plants that have been the focus of ethnobotanical surveys of the Micronesian islands of Palau and Pohnpei and compiled in Primary Health Care Manuals (32, 33). The plants from these Micronesian islands are known for botanical endemism, and are the focus of an NIH-funded project to develop computational methods for identifying and validating potential therapeutic knowledge about plants. A co-occurrence based metric was use to rank the plant-associated disease concepts from both MEDLINE and the Primary Health Care Manuals for Palau and Pohnpei. The identified associations from these two sources were compared to highlight known ethnobotanical uses that have been evaluated in indexed biomedical literature as well as uses that may be unique. The evaluation for this study focused on potential therapeutic uses of plants for the ICD-9-CM *Mental, Behavioral and Neurodevelopmental Disorders* category. This specific focus was chosen as such conditions may be related to stress-related suicide, for which Micronesia has amongst the highest rates in the world (Department of Public Health and Social Security) (32). The results, which included the identification of several putative therapeutic uses of plant species, reveal the applicability of informatics approaches for supporting large-scale comparative analysis of ethnobotanical knowledge within biomedical literature. The results also suggest that indigenous knowledge may be used to guide the identification of bioactive plant species within a western medicinal context.

Methods

The goal of this study was to identify disease associations of plants identified within the Primary Health Care Manuals (PHCMs) of Palau and Pohnpei (32, 33). Processing articles related to these plants as identified from MEDLINE using PubMed provided the basis for comparison of identified associations from the Primary Health Care Manuals. A general overview of the process developed for this study is depicted in Figure 1.

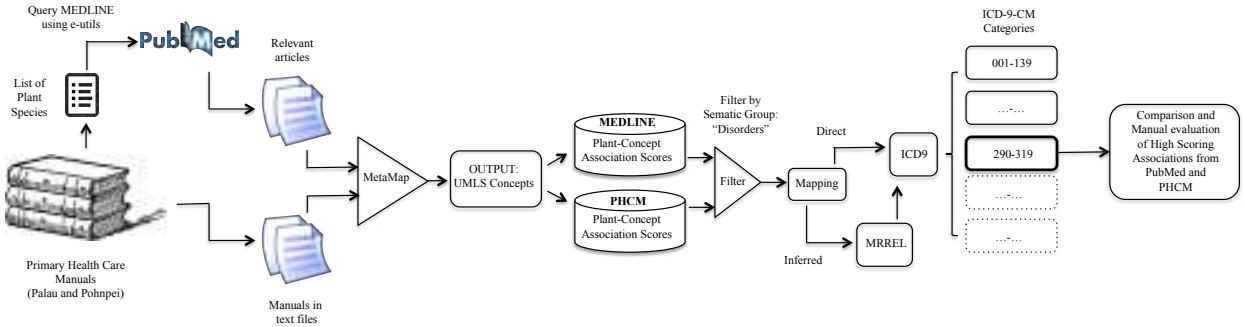

Figure 1: *Study Overview.* MEDLINE citations accessed from PubMed containing plants from the Primary Health Care Manuals (PHCMs) for Palau and Pohnpei and corresponding descriptions in the PHCMs were processed using MetaMap. Plant associations with UMLS identified concepts were scored and filtered by the "Disorders" UMLS Semantic Group. Manual evaluation focused on disease concepts that could be mapped to the ICD-9-CM category *"Mental, Behavioral And Neurodevelopmental Disorders* (290-319)*."*

Biomedical concept association with plants from Micronesia. A list of plants was attained from the Primary Health Care Manuals (PHCMs) of Palau (32) and Pohnpei (33), which were the result of ethnobotanical explorations carried out in the Micronesian Islands of Palau and Pohnpei. A Ruby script that leveraged Entrez e-utils was used to query MEDLINE from PubMed with each plant's scientific name. The titles and abstracts of the resulting set of identified MEDLINE citations were extracted using e-utils, and processed using the MetaMap Java API (34). The Unified Medical Language System (UMLS) concepts identified by MetaMap were parsed from the machine output and the association scores between a given plant and UMLS concept calculated using the following equation:

$$Score(p,c) = f_c \times \log\frac{N}{n_p} \qquad (1)$$

where, f_c is the frequency of concept 'c' co-occurring with plant 'p', N is the total number of plants and n_p is the number of plants co-occurring with the given concept 'c'. These scores were then normalized using the following formula:

$$nS(p, c_i) = \frac{Score(p,c_i)}{\max(Score(p,c))} \qquad (2)$$

where, $S(p, c_i)$ is the score of a given plant-concept association and *max(Score(p,c))* is the maximum score among all the concepts associated with a given plant *p*.

The descriptions from the PHCMs, which consisted of one plant per document, were processed using MetaMap and the resulting plant-concept associations were scored as with MEDLINE articles using Equations 1 and 2.

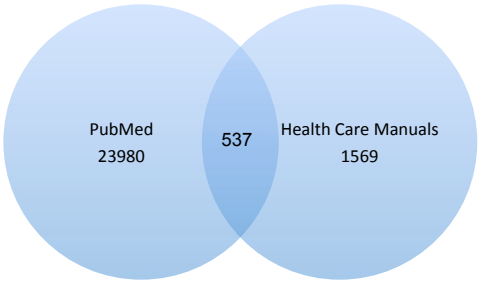

Figure 2: *Identified plant and "Disorder" UMLS Semantic Group concept associations.*

Mapping to ICD-9-CM categories. Concepts were filtered based on those that belonged to the "Disorders" UMLS Semantic Group and mapped to corresponding ICD-9-CM codes by querying the UMLS MRCONSO table (these were referred to as "direct mappings"). If no direct mapping was possible for a given concept, related concepts were retrieved from UMLS MRREL table (including all relationship types) and then mapped to ICD-9-CM (these were referred to as "inferred mappings").

Evaluation. The validity of the results focused on those plant associations with concepts that belonged the ICD-9-CM category "*Mental, Behavioral And Neurodevelopmental Disorders* (290-319)." The evaluation involved manually comparing the MetaMap-identified top scoring plant-concept associations from MEDLINE with those identified from the PHCMs.

Results

Plant-concept associations. From the PHCMs a total of 180 unique plants were identified, for which 129 could be identified in MEDLINE. In total there were 19,798 citations identified from MEDLINE that had at least one mention of a plant from the PHCMs, out of which 18,322 contained associated UMLS concepts. A total of 22,425 and 310,155 plant-concept co-occurrences were identified from the PHCMs and MEDLINE datasets, respectively, with 7,521 associations in common. The organizing and filtering of associations based on the semantic group '*Disorders*' resulted in 2,106 and 24,517 associations for PHCMs and MEDLINE datasets, respectively, with 537 associations in common (Figure 2). Comparative statistics of plant-wise associations across different semantic types belonging to the semantic group '*Disorders*' is presented in Supplemental Table 1. The score comparison of plant-concept associations across different semantic types of semantic group '*Disorders*' is presented in Supplemental Table 2. Similar statistics for the ICD-9-CM categories are provided in Supplemental Table 3 and 4.

Distribution across ICD-9-CM disease categories. The plant-associated disorder concepts were mapped into 17 ICD-9-CM categories (Table 1). Out of total 526 concepts from the PHCMs dataset, 337 (64%) were mapped to ICD-9-CM codes (103 direct and 234 inferred). A total of 3,189 out of 5,393 concepts (59%) from the MEDLINE dataset were mapped to ICD-9-CM codes (786 direct and 2403 inferred). Associations from the MEDLINE dataset were represented in all 17 categories, while those from PHCMs dataset were represented in 16 categories (all except "*Certain Conditions Originating In The Perinatal Period*" [760-779]).

Evaluation. There were 27 disorder concept associations for 22 plants identified from the PHCMs dataset belonging to ICD-9-CM category "*Mental, Behavioral And Neurodevelopmental Disorders* (290-319)," comprised of seven unique UMLS concepts: Stress (C0038435), Acute Psychosis (C0281774), Hand Rubbing (C0239846), Stupor (C0085628), Depression (C0011570), Depressive disorder (C0011581), and Anxiety disorder (C0003469). By contrast, there were 635 disease concept associations from MEDLINE belonging to the same ICD-9-CM category comprised of 145 UMLS concepts. Table 2 summarizes the number of associations manually identified as correct. For the selected set of plants, there were eight true associations out of 14 top scoring associations from MEDLINE dataset across all disease categories. Sixteen out of the 22 top scoring associations from PHCM dataset were also deemed to be valid across all disease categories. Seven out of 11 and 10 out of 22 top scoring associations from the "290-319" category were true for MEDLINE and PHCM dataset, respectively. Table 3 summarizes the manual evaluation of the top scoring predicted therapeutic use for MetaMap identified disease concepts from the PHCMs

and MEDLINE. Within the chosen ICD-9-CM category, six plants were only identified in the PHCMs with therapeutic use but not in MEDLINE: *Glochidion ramiflorum, Horsfieldia irya, Phaleria nisidai, Calophyllum innophyllum,* and *Phyllanthus palauensis.* There were 111 plants identified in MEDLINE in categories that did not have corresponding PHCM association for the chosen ICD-9-CM category.

Table 1: ***Distribution of Plant-Human Disease Concept Associations Across ICD-9-CM Categories.*** The disease concept counts include all associations that could be mapped to ICD-9-CM codes (either through direct or inferred mapping).

Code	Category	MEDLINE	Both	PHCM
001-139	Infectious and parasitic diseases	1321	22	109
140-239	Neoplasms	1676	26	110
240-279	Endocrine, nutritional and metabolic diseases, and immunity disorders	1126	2	19
280-289	Diseases of the blood and blood-forming organs	353	4	34
290-319	Mental, behavioral and neurodevelopmental disorders	635	3	27
320-389	Diseases of the nervous system and sense organs	1193	23	109
390-459	Diseases of the circulatory system	651	8	32
460-519	Diseases of the respiratory system	259	2	21
520-579	Diseases of the digestive system	993	25	81
580-629	Diseases of the genitourinary system	506	6	30
630-679	Complications of pregnancy, childbirth, and the puerperium	124	0	4
680-709	Diseases of the skin and subcutaneous tissue	463	10	80
710-739	Diseases of the musculoskeletal system and connective tissue	226	1	27
740-759	Congenital anomalies	507	1	8
760-779	Certain conditions originating in the perinatal period	92	0	0
780-799	Symptoms, signs, and ill-defined conditions	1811	69	294
800-999	Injury and poisoning	809	65	200

Discussion

Along with continued progress of ethnobotanical surveys of indigenous and other populations around the globe, there are increasing efforts to digitize historical texts (35) and other documentation of traditional knowledge that may include descriptions of therapeutic applications of plant species (36). Informatics methodologies may be used to connect such cultural knowledge that has remained historically isolated from contemporary biomedical knowledge sources, such as biomedical literature. Information regarding the historical use of plants may potentially reflect the efficacy and safety of their use. The current regulatory guidelines established by the Center for Drug Evaluation and Research (CDER) for botanicals encourages submission of documentation of prior human experience for preliminary safety assessments (37). Such information may also provide relevant background for conducting search for chemical drugs as well as in designing appropriate clinical studies for evaluation (38). This study aimed to develop an informatics approach for enabling comparison and evaluation of potential therapeutic information documented in PHCMs of Palau and Pohnpei in light of potentially supporting evidence within MEDLINE.

Table 2: Counts of True Associations Identified from MEDLINE and PHCMs

	All ICD-9-CM Disease Categories	ICD-9-CM 290-319 Category
MEDLINE	8/14	7/11
PHCMs	16/22	10/22

Table 3: Manual Assessment of Top Identified Disease Concept Associations for 22 Plants from the PHCMs and MEDLINE the ICD-9-CM category "Mental, Behavioral And Neurodevelopmental Disorders (290-319)." Shown for each top hit are the Disease Concept, UMLS CUI, ICD Category (only for associations across All ICD-9-CM Categories), Rank Score, and Manual Evaluation (True/False) along with citation used for evaluation.

	MEDLINE[1]		**PHCM[2]**	
	All ICD-9-CM Categories	**290-319 ICD-9-CM Category**	**All ICD-9-CM Categories**	**290-319 ICD-9-CM Category**
Ageratum conyzoides L.	Vein Disorder (C0235522) (390-459) Score: 0.0833 False [PMID: 10544139]	Impotence (C0242350) Score: 0.0130 False [PMID: 17362507]	Pallor (C0030232) (780-799) Score: 0.5531 False [Pl:pp 37]	Stress (C0038435) Score: 0.1225 False [Pl:pp 37]
Allophylus timoriensis (DC.) Blume	No association identified	No association identified	Precursor T-Cell Lymphoblastic Leukemia-Lymphoma (C1961099) (280-289) Score: 0.3559 False [Ph:pp 32]	Stress (C0038435) Score: 0.0063 True [Ph:pp 106]
Areca catechu L.	Fibroses, Oral Submucous (C0029172) (520-579) Score: 0.0909 False [PMID: 26336810]	Schizophrenias (C0036341) Score: 0.0682 True [PMID 19748131]	Rubor (C0332575) (780-799) Score: 0.6111 False [Ph:pp 121]	Acute Psychosis (C0281774) Score: 0.0208 False [Ph:pp 121]
Calophyllum innophyllum L.	No association identified	No association identified	Carcinogenesis (C0596263) (140-239) Score: 0.8182 True [Pl:pp 103]	Stress (C0038435) Score: 0.0221 True [Pl:pp 103]
Centella asiatica (L.) Urb.	Oxidative Stress (C0242606) (760-779) Score: 0.0544 True [PMID: 25633675]	Anxiety Disorder (C0003469) Score: 0.0259 True [PMID: 22841896]	Dysentery (C0013369) (001-139) Score: 0.8811 True [Ph:pp 30]	Hand Rubbing (C0239846) Score: 0.1006 False [Ph:pp 30]
Citrus limon (L.) Burm. f.	Facial dysmorphism, immunodeficiency, livedo, and short stature (C3554576) (240-279) Score: 0.0521 False [PMID: 12231681]	Stress (C0038435) Score: 0.0159 True [PMID 26050208]	Tachycardia (C0039231) (780-799) Score: 0.1539 True [Ph:pp 117]	Stupor (C0085628) Score: 0.0769 True [Ph:pp 117]
Clerodendrum inerme (L.) Gaertn	Tic Dis Motor (C0751554) (290-319) Score: 0.1231 True [PMID: 19617461]	Tic Dis Motor (C0751554) Score: 0.1231 True [PMID: 19617461]	Asthma (C0004096) (460-519) Score: 1.0000 True [Ph:pp 99]	Stress (C0038435) Score: 0.0210 True [Ph:pp 99]
Cyathula prostrata (L.) Blume	Ascites (C0003962) (780-799) Score: 0.27017 True [PMID: 23870465]	No association identified	Thin hair (C0423867) (680-709) Score: 0.3428 False [Ph:pp 10]	Hand Rubbing (C0239846) Score: 0.0571 False [Ph:pp 70]
Glochidion ramiflorum J. R. Forst. & G. Forst.	No association identified	No association identified	Joint Pain Adverse Event (C1963066) (710-739) Score: 0.2338 Arthralgia (C0003862) (710-739) Score: 0.2338 True [Ph:pp 150]	Stress (C0038435) Score: 0.0284 True [Ph:pp 102]
Hibiscus tiliaceus L.	Disorder, Puerperal (C0034040) (580-629) Score: 0.0608 True [PMID: 22494845]	Disease, Seitelberger's (C0270724) Score: 0.0308 False [PMID: 16701930]	Diarrhea (C0011991) (780-799) Score: 0.1961 True Ph:pp 27	Stress (C0038435) Score: 0.1218 True [Ph:pp 99]

Horsfieldia irya (Gaertn.) Warb	No association identified	No association identified	Fainting (C0039070) (780-799) Score: 0.2802 True [Pl:pp 104]	Stress (C0038435) Score: 0.0682 True [Pl:pp 104]
Ipomoea littoralis Blume	No association identified	No association identified	Staphylococcus aureus infections (C1318973) (001-139) Score: 0.2498 True [Ph:pp 45]	Hand Rubbing (C0239846) Score: 0.0555 False [Ph:pp 142]
Ipomoea mauritiana Jacq.	Diastasis (C0036679) (320-389) Score: 0.1155 False [PMID: 25050305]	No association identified	Muscle Cramps (C0026821) (320-389) Score: 0.5835 True [Ph:pp 133]	Hand Rubbing (C0239846) Score: 0.1110 False [Ph:pp 133]
Ixora casei Hance	Drug Tolerance (C0013220) (290-319) Score: 0.3393 False [PMID: 19283052]	Drug Tolerance (C0013220) Score: 0.3393 False [PMID: 19283052]	Dysmenorrhea (C0013390) (580-629) Score: 0.1805 True [Pl:pp 135]	Hand Rubbing (C0239846) Score: 0.0333 False [Ph:pp 39]
Kyllinga brevifolia Rottb	Catatonia (C0007398) (780-799) Score: 0.2500 True [PMID: 10473172]	Stress (C0038435) Score: 0.0509 True [PMID: 10473172]	Viral infection (C0042769) (780-799) Score: 1.0000 True [Ph:pp 49]	Hand Rubbing (C0239846) Score: 0.0238 False [Ph:pp 69]
Nephrolepis obliterata (R. Br.) J. Sm.	NA	NA	C0013369: Dysentery (001-139) Score: 0.3231 True Ph:pp 31	C0239846: Hand Rubbing Score: 0.0369 False [Ph:pp 30]
Paraderris elliptica (Wall.) Adema	IGS (C1306856) (520-579) Score: 0.6688 False [PMID: 23144360]	No association identified	Conjunctivitides (C0009763) (320-389) Score: 0.8182 True [Pl:pp 62]	Depressions (C0011570) Score: 0.0584 False [Pl:pp 62]
Phaleria nisidai Kanehira	Fallot Tetralogy (C0039685) (740-759) Score: 0.19448 False [PMID: 23144360]	No association identified	Tumor Mass (C3273930) (140-239) Score: 0.7007 True [Pl:pp 105]	Stress (C0038435) Score: 0.0284 True [Pl:pp 105]
Phyllanthus palauensis Hosok.	No association identified	No association identified	Flushing (C0016382) (780-799) Score: 0.5000 False [Pl:pp 105]	Stress (C0038435) Score: 0.0406 True [Pl:pp 105]
Piper methysticum G. Forst.	Anxiety Disorder (C0003469) (290-319) Score: 0.1796 True [PMID: 23635869]	Anxiety Disorder (C0003469) Score: 0.1796 True [PMID: 23635869]	Skin Discoloration (C0151907) (680-709) Score: 0.1250 True Ph:pp 105 [toxicity]	Anxiety Disorder (C0003469) Score: 0.0417 True [Ph:pp 105]
Premna serratifolia L.	Myopathy (C0026848) (320-389) Score: 0.1114 True [PMID: 23407688]	Stress (C0038435) Score: 0.0339 True [PMID: 23244417]	Nodule (C0028259) (140-239) Score: 0.4927 False [Ph:pp 51]	Hand Rubbing (C0239846) Score: 0.0035 False [Ph:pp 51]
Solenostemon scutellarioides (L.)	Prostration (C0277794) (320-389) Score: 0.1175 False [PMID: 18603655]	Stress (C0038435) (290-319) Score: 0.0239 False [PMID: 2586264]	Rash (C0015230) (780-799) Score: 1.0000 True [Ph:pp 55]	Depressions (C0011570) Score: 0.1922 False [Ph:pp 55]

[1]Evaluation based on PHCMs, page numbers in brackets (Pl = PHCM of Palau; Ph = PHCM of Pohnpei); [2]Evalation based on biomedical literature, PubMed identifier in brackets

From this study, several potentially interesting therapeutic applications were highlighted based on the scoring metric (Supplemental Tables 1-4). This scoring metric was inspired by the *tf-idf* (term frequency-inverse document frequency) metric. An advantage of using the *tf-idf* weighing strategy is that it is able to highlight rare interactions and filter noisy datasets by deprioritizing trivial relationships. Future work may involve additional statistical comparisons of additional scoring metrics (e.g., Mutual Information Measure, Association rules). The identified concepts were organized according to the UMLS Semantic Network, which reduces the more than 1 million concepts in the UMLS Metathesaurus into 133 semantic types (39) that are further grouped into 15 semantic groups (40). Using the UMLS Semantic Network helped focus the study by grossly filtering identified associations into a single semantic group, "*Disorders*." It is important to note, however, that it is often challenging to correlate the diagnosis descriptions from indigenous or historical sources into canonical medical concepts (37). For example, within the *Herbarium Ambionense* (41), a compilation published in 1741 about plants from Amboina (a geographic region within modern day Indonesia) gonorrhea is described as "fire-piss."

The International Classification of Diseases (ICD) provides a set of standardized codes for describing diagnosis maintained by World Health Organization (42). ICD-9 Clinical Modification (ICD-9-CM) is the adaptation of the ninth revision of ICD created by the U.S. National Center for Health Statistics (NCHS), Center for Disease Control (CDC) (43). By mapping disease concepts to the ICD classification, a broader picture of potential therapeutic applications of plants was possible in this study. For the purposes of this feasibility study, the evaluation of identified potential therapies was limited to the ICD-9-CM category "*Mental, Behavioral And Neurodevelopmental Disorders* (290-319)." Six out of 22 plant species from Palau and Pohnpei showed promising medicinal applications for the ICD-9-CM evaluation category based on MEDLINE indexed literature beyond the original applications described in the Primary Health Care Manuals. Association between *Areca catechu* (betel nut) and schizophrenia or schizoaffective disorder was highlighted by the approach implemented in this study. Published evidence has shown that high-consumption of betel nuts had significantly milder positive symptoms of schizophrenia in males (44, 45). Such a therapeutic role of betel nut has been attributed to arecoline, a partial muscarinic agonist (46). Its potential use in medication-induced movement disorders (46) and role in enhancing cognitive ability and social activity in schizophrenia patients have also been explored (47). The use of *Citrus limon* essential oil has been evaluated and has shown positive effects for anxiety (48-50). The possibility of action on benzodiazepine-type receptors is implicated for such anxiolytic activity (48). *Clerodendrum inerme*, traditionally used to reduce stress, has been successfully tested for use against motor tics (*Tic disorder*) (51). The therapeutic role of hispidulin, a flavonoid isolated from *C. inerme* has been proposed in hyper-dopaminergic disorders (52). The role of *Kyllinga brevifolia* in alleviating stress has been implicated by interaction with benzodiazepine receptors (53). *Premna serratifolia* has been shown to play a role in stress resistance. An iridoid, 10-O-trans-p-Coumaroylcatalpol, from *P. serratifolia* decreased the aggregation of Parkinson's disease associated protein (alpha synuclein) in a transgenic *Caenorhabditis elegans* model, as well as promoting longevity (54). The anxiolytic effect of *Centella asiatica* extract has been demonstrated in mice, with activity attributed to madecassoside and asiaticoside (55). Additionally, has been shown to attenuate amyloid-beta induced oxidative stress and mitochondrial dysfunction contributing towards its neuroprotective action (56). In addition to the role of identified plants species discussed here for the ICD-9-CM category "*Mental, Behavioral And Neurodevelopmental Disorders* (290-319)," several other potential therapeutic roles were identified (Table 2). The ability to identify such potential uses of plants that have been identified from study of indigenous populations may provide the basis for designing appropriate subsequent clinical studies to understand the efficacy of these plant species.

This study aimed to extract and prioritize the associations between plants and biomedical concepts based on co-occurrence statistics. A major weakness of the current co-occurrence approach is that cannot robustly distinguish between correlation and causation relationships. Thus, while the developed approach did highlight some potentially useful therapeutic associations, it did not distinguish between therapeutic and toxicity associations. For example, in addition to the above mentioned potential therapeutic roles, *Areca catechu* is also known to cause Oral Submucous Fibrosis (57) and oral cancer (58). Further confounding this issue is the fact that plant toxicity profiles may sometimes be used to identify additional therapeutic roles. For example, the toxic plant *Atropa belladonna* has been shown to be a potential analgesic (59). A major area of future emphasis of this work will thus be enhancement of the approaches demonstrated in this study for distinguishing between toxicity and therapeutic roles. An immediate area of emphasis will be to focus on mining associations within context such as within defined scopes of text (e.g., sentence, phrase, or utterance). It will be essential to implement such enhancements before considering the general utility of informatics to support the ethnobotany workflow towards identification of possible plant-based therapies. Nonetheless, the results of this feasibility study demonstrate the potential to enhance the current workflow in identifying previously described therapeutic uses for plants.

Conclusion

The adoption of informatics for identifying ethnobotanical knowledge within the increasing stores of electronic has been slow. This feasibility study demonstrates the potential to leverage existing natural language processing, along with terminological resources, to mine medicinal plant knowledge from biomedical literature as indexed in MEDLINE. The developed approach was able to identify potential new therapeutic applications for plants previously described in ethnobotanical surveys from the Micronesian islands of Palau and Pohnpei. Building on this feasibility study, future work is needed to increase overall specificity of identified correlations (e.g., distinguish between reported therapeutic versus toxic effects) before computational approaches be used for potentially impacting the ability to identify new plant based remedies inspired by ethnobotanical surveys.

Supplementary information referenced in this study is available at:
https://sites.google.com/a/brown.edu/phytokb/tbi2016

Acknowledgements

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. The cited Primary Health Care Manuals from Palau and Pohnpei are a result from a collaborative program involving The Nature Conservancy—Micronesia Office, The Conservation Society of Pohnpei (CSP), the Federated States of Micronesia Ministry of Health, Pohnpei Council of Traditional Leaders, the Pohnpei State Department of Health Services, the Pohnpei State Government, Pohnpei State Hospital, Pohnpei Department of Lands and Natural Resources, Pohnpei Department of Economic Affairs, Island Food Community of Pohnpei, College of Micronesia—FSM, the Belau National Museum, The Ministry of Health, Republic of Palau, the Continuum Center for Health and Healing at Beth Israel Medical Center—NY, The National Tropical Botanical Garden—HI, and The New York Botanical Garden. The authors from the New York Botanical Garden (WL & MJB) are very grateful to the V. Kann Rasmussen Foundation, the Gildea Foundation, the MetLife Foundation, the Overbrook Foundation, Edward P. Bass and the Philecology Trust, the Prospect Hill Foundation, the Marisla Foundation, and The Germeshausen Foundation for their support of our field studies in Micronesia.

References

1. Balick MJ, Cox PA. Plants, People, and Culture: The Science of Ethnobotany: Scientific American Library; 1997.
2. Noble RL. The discovery of the vinca alkaloids--chemotherapeutic agents against cancer. Biochem Cell Biol. 1990;68(12):1344-51.
3. Balick MJ. Ethnobotany and the identification of therapeutic agents from the rainforest. Bioactive compounds from plants. 1990;154:22-39.
4. NIH. Citations Added to MEDLINE by Fiscal Year 2015. Available from: http://www.nlm.nih.gov/bsd/stats/cit_added.html.
5. Sharma V, Sarkar IN. Bioinformatics opportunities for identification and study of medicinal plants. Briefings in bioinformatics. 2013;14(2):238-50.
6. Cotton CM. Ethnobotany: principles and applications: John Wiley & Sons; 1996.
7. Bennett BC, Balick MJ. Does the name really matter? The importance of botanical nomenclature and plant taxonomy in biomedical research. Journal of ethnopharmacology. 2014;152(3):387-92.
8. Schultes RE, Reis Sv. Ethnobotany: evolution of a discipline: Chapman & Hall Ltd; 1995.
9. Saslis-Lagoudakis CH, Klitgaard BB, Forest F, Francis L, Savolainen V, Williamson EM, et al. The use of phylogeny to interpret cross-cultural patterns in plant use and guide medicinal plant discovery: an example from Pterocarpus (Leguminosae). PloS one. 2011;6(7):e22275.
10. Thomas MB. Emerging synergies between information technology and applied ethnobotanical research. 2003.
11. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349(6245):261-6.
12. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:17-21.
13. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on translational bioinformatics. 2009;2009:56-60.
14. Hanauer DA, Saeed M, Zheng K, Mei Q, Shedden K, Aronson AR, et al. Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: a feasibility analysis. J Am Med Inform Assoc. 2014;21(5):925-37.

15. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics. 2009;10 Suppl 9:S14.
16. Sharma V, Sarkar IN. Leveraging concept-based approaches to identify potential phyto-therapies. Journal of biomedical informatics. 2013;46(4):602-14.
17. Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H. Ranking gene-drug relationships in biomedical literature using Latent Dirichlet Allocation. Pac Symp Biocomput. 2012:422-33.
18. Weeber M, Klein H, de Jong-van den Berg LTW, Vos R. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. J Am Soc Inf Sci Tec. 2001;52(7):548-57.
19. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inform Assoc. 2003;10(3):252-9.
20. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol. 2010;6(9).
21. Law J, Bauin S, Courtial JP, Whittaker J. Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. Scientometrics. 1988;14(3-4):251-64.
22. Abacha AB, Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomedical Semantics. 2011;2(S-5):S4.
23. Ben Abacha A, Zweigenbaum P. A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts. In: Gelbukh A, editor. Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science. 6609: Springer Berlin Heidelberg; 2011. p. 139-50.
24. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008;15(1):87-98.
25. de Bruijn B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inform. 2002;67(1-3):7-18.
26. Hirschman L, Park JC, Tsujii J, Wong L, Wu CH. Accomplishments and challenges in literature data mining for biology. Bioinformatics. 2002;18(12):1553-61.
27. Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. Medinfo. 2004;11(Pt 2):758-62.
28. Srinivasan P, Rindflesch T. Exploring text mining from MEDLINE. Proceedings / AMIA Annual Symposium AMIA Symposium. 2002:722-6.
29. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc. 2006:349-53.
30. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Briefings in bioinformatics. 2007;8(5):358-75.
31. Ball R, Botsis T. Can Network Analysis Improve Pattern Recognition Among Adverse Events Following Immunization Reported to VAERS? Clinical Pharmacology & Therapeutics. 2011;90(2):271-8.
32. Dahmer SM BM, Kitalong AH, Kitalong C, Herrera K, Law W, Lee R, Tadao V, Rehuher F, Hanser S, Soaladaob K, Ngirchobong G, Besebes M, Wasisang F, Kulakowski D, Adam I Palau Primary Health Care Manual: Health Care in Palau: Combining Conventional Treatments and Traditional Uses of Plants for Health and Healing: CreateSpace Independent Publishing Platform; 1 edition (May 15, 2012); 2012.
33. Lee R SN, Balick MJ, Sohl F, Roberts AS, Herrera K, Dahmer S, Lieskovsky M, Raynor W, Raynor P, Albert E, Trauernicht C, Offringa L, Adam I, Law W, Hunt M (Contributor), Alfred Dores A (Contributor). Pohnpei Primary Health Manual: Health Care in Pohnpei, Micronesia: Traditional Uses of plants for Health and Healing: CreateSpace Independent Publishing Platform (August 17, 1010); 2010.
34. MetaMap Java API. Available from: http://metamap.nlm.nih.gov/JavaApi.shtml.
35. Biodiversity Heritage Library. Available from: http://biodiversitylibrary.org.
36. The Micronesia Challenge. Available from: http://www.micronesiachallenge.org/.
37. Chen ST, Dou J, Temple R, Agarwal R, Wu KM, Walker S. New therapies from old medicines. Nat Biotechnol. 2008;26(10):1077-83.
38. Buenz EJ, Schnepple DJ, Bauer BA, Elkin PL, Riddle JM, Motley TJ. Techniques: Bioprospecting historical herbal texts by hunting for new leads in old tomes. Trends in pharmacological sciences. 2004;25(9):494-8.
39. McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003;4(1):80-4.
40. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform. 2001;84(Pt 1):216-20.
41. Rumphius G. Herbarium Ambionense, Johannes Burman. 1741-1755.

42. WHO. World Health Organization. Available from: http://www.who.int/classifications/icd/en/.
43. NCHS. National Center for Health Statistics, Center for Disease Control (CDC). Available from: http://www.cdc.gov/nchs/icd/icd9cm.htm.
44. Sullivan RJ, Allen JS, Otto C, Tiobech J, Nero K. Effects of chewing betel nut (Areca catechu) on the symptoms of people with schizophrenia in Palau, Micronesia. Br J Psychiatry. 2000;177:174-8.
45. Sullivan RJ, Andres S, Otto C, Miles W, Kydd R. The effects of an indigenous muscarinic drug, Betel nut (Areca catechu), on the symptoms of schizophrenia: a longitudinal study in Palau, Micronesia. Am J Psychiatry. 2007;164(4):670-3.
46. Bales A, Peterson MJ, Ojha S, Upadhaya K, Adhikari B, Barrett B. Associations between betel nut (Areca catechu) and symptoms of schizophrenia among patients in Nepal: A longitudinal study. Psychiatry Res. 2009;169(3):203-11.
47. Adilijiang A, Guan T, He J, Hartle K, Wang W, Li X. The Protective Effects of Areca catechu Extract on Cognition and Social Interaction Deficits in a Cuprizone-Induced Demyelination Model. Evid Based Complement Alternat Med. 2015;2015:426092.
48. C LML, Goncalves e Sa C, de Almeida AA, da Costa JP, Marques TH, Feitosa CM, et al. Sedative, anxiolytic and antidepressant activities of Citrus limon (Burn) essential oil in mice. Pharmazie. 2011;66(8):623-7.
49. Khan RA, Riaz A. Behavioral effects of Citrus limon in rats. Metab Brain Dis. 2015;30(2):589-96.
50. Setzer WN. Essential oils and anxiolytic aromatherapy. Nat Prod Commun. 2009;4(9):1305-16.
51. Fan PC, Huang WJ, Chiou LC. Intractable chronic motor tics dramatically respond to Clerodendrum inerme (L) Gaertn. J Child Neurol. 2009;24(7):887-90.
52. Huang WJ, Lee HJ, Chen HL, Fan PC, Ku YL, Chiou LC. Hispidulin, a constituent of Clerodendrum inerme that remitted motor tics, alleviated methamphetamine-induced hyperlocomotion without motor impairment in mice. Journal of ethnopharmacology. 2015;166:18-22.
53. Hellion-Ibarrola MC, Ibarrola DA, Montalbetti Y, Villalba D, Heinichen O, Ferro EA. Acute toxicity and general pharmacological effect on central nervous system of the crude rhizome extract of Kyllinga brevifolia Rottb. Journal of ethnopharmacology. 1999;66(3):271-6.
54. Shukla V, Phulara SC, Yadav D, Tiwari S, Kaur S, Gupta MM, et al. Iridoid compound 10-O-trans-p-coumaroylcatalpol extends longevity and reduces alpha synuclein aggregation in Caenorhabditis elegans. CNS Neurol Disord Drug Targets. 2012;11(8):984-92.
55. Wanasuntronwong A, Tantisira MH, Tantisira B, Watanabe H. Anxiolytic effects of standardized extract of Centella asiatica (ECa 233) after chronic immobilization stress in mice. Journal of ethnopharmacology. 2012;143(2):579-85.
56. Gray NE, Sampath H, Zweig JA, Quinn JF, Soumyanath A. Centella asiatica Attenuates Amyloid-beta-Induced Oxidative Stress and Mitochondrial Dysfunction. J Alzheimers Dis. 2015;45(3):933-46.
57. Pant I, Kumar N, Khan I, Rao SG, Kondaiah P. Role of Areca Nut Induced TGF-beta and Epithelial-Mesenchymal Interaction in the Pathogenesis of Oral Submucous Fibrosis. PloS one. 2015;10(6):e0129252.
58. Li YC, Chang JT, Chiu C, Lu YC, Li YL, Chiang CH, et al. Areca nut contributes to oral malignancy through facilitating the conversion of cancer stem cells. Mol Carcinog. 2015.
59. Owais F, Anwar S, Saeed F, Muhammad S, Ishtiaque S, Mohiuddin O. Analgesic, Anti-inflammatory and neuropharmacological effects of Atropa belladonna. Pak J Pharm Sci. 2014;27(6 Spec No.):2183-7.