Leveraging biodiversity knowledge for potential phyto-therapeutic applications

Vivekanand Sharma,[1] Indra Neil Sarkar[1,2,3]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001445).

[1]Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont, USA
[2]Biomedical Informatics Unit, Center for Clinical and Translational Science, University of Vermont, Burlington, Vermont, USA
[3]Department of Computer Science, University of Vermont, Burlington, Vermont, USA

Correspondence to
Dr Indra Neil Sarkar, Center for Clinical and Translational Science, University of Vermont, 89 Beaumont Avenue, Given Courtyard N309, Burlington, VT 05405, USA; neil.sarkar@uvm.edu

Received 26 October 2012
Revised 2 February 2013
Accepted 2 March 2013
Published Online First 21 March 2013

ABSTRACT

Objective To identify and highlight the feasibility, challenges, and advantages of providing a cross-domain pipeline that can link relevant biodiversity information for phyto-therapeutic assessment.

Materials and methods A public repository of clinical trials information (ClinicalTrials.gov) was explored to determine the state of plant-based interventions under investigation.

Results The results showed that ~15% of drug interventions in ClinicalTrials.gov were potentially plant related, with about 60% of them clustered within 10 taxonomic families. Further analysis of these plant-based interventions identified ~3.7% of associated plant species as endangered as determined from the International Union for the Conservation of Nature Red List.

Discussion The diversity of the plant kingdom has provided human civilization with life-sustaining food and medicine for centuries. There has been renewed interest in the investigation of botanicals as sources of new drugs, building on traditional knowledge about plant-based medicines. However, data about the plant-based biodiversity potential for therapeutics (eg, based on genetic or chemical information) are generally scattered across a range of sources and isolated from contemporary pharmacological resources. This study explored the potential to bridge biodiversity and biomedical knowledge sources.

Conclusions The findings from this feasibility study suggest that there is an opportunity for developing plant-based drugs and further highlight taxonomic relationships between plants that may be rich sources for bioprospecting.

BACKGROUND AND SIGNIFICANCE

Biodiversity provides the foundation for human health and well-being by providing the basic requirements of life. The biotic diversity in genetic and biochemical components has been harnessed to secure life-sustaining food and medicinal sources. Recent years have seen the advent of advanced approaches for analyzing molecular and genetic details across the spectrum of life, including plants. This has resulted in the cataloguing of vast amounts of potentially insightful plant-specific knowledge in publications or curated databases. With increasing awareness about the importance of biodiversity, informatics approaches are being developed that aim at augmenting traditional bioinformatics techniques by encompassing a wider spectrum of data types and organisms.[1] This research area, termed 'biodiversity informatics,' incorporates informatics principles to accommodate the full range of biological information, from molecules to populations.[2] Approaches that focus on unifying biodiversity information for use in other domains (eg, biomedicine) have highlighted and leveraged the species-centric nature of this discipline.[1 3] For example, named entity recognition tools designed to identify organism scientific names have been used alongside biomedical ontology-based annotations to link organism-specific information across resources like GenBank.[4 5] Such approaches are essential for navigating across the domains of contemporary and archival information. For example, for medicinal applications of plants, the linking of legacy data (eg, ethnobotanical) with contemporary biomedical and clinical data can promote prioritization of conservation strategies for species of medicinal interest as well as developing bioprospecting strategies.[6–8] Breaking the barriers and enabling bidirectional flow of information between the biodiversity and biomedical domains may enable the harnessing of biodiversity knowledge for pharmaceutical leads.

This study aimed to explore the potential for developing an approach for linking the generally unlinked domains of biodiversity and biomedical knowledge. Specifically, an informatics pipeline was developed for the extraction and integration of data related to plant species along with data available about clinical trials involving plant species. The resulting workflow demonstrates the potential for linking canonically biodiversity and biomedical data sources and thus the possibility of leveraging biodiversity knowledge for biomedical applications, such as the identification of potential phyto-therapies.

A CASE STUDY FOR LINKING BIODIVERSITY AND BIOMEDICAL KNOWLEDGE: MEDICINAL PLANTS

Plant-based medicines have been used for ages, generally based on folklore passed down from generation to generation. In contemporary medicine, study of such knowledge has been used to identify bioactive plant metabolites with therapeutic importance. Several commercially important drugs have been isolated or developed that take advantage of plant biodiversity (eg, diosgenin from *Dioscorea nipponica*[9] and analgesic aspirin from willow bark (*Salix* sp)[10]). Recently, there has been an increasing interest in discovering plant-based drugs.[11] However, the process of discovering a plant-based drug has numerous challenges. Amidst the plethora of historical or ethnobotanical texts describing medicinal applications of plants, it remains challenging to verify such descriptions in light of contemporary scientific methodologies and regulations. Furthermore, the generally isolated and

To cite: Sharma V, Sarkar IN. *J Am Med Inform Assoc* 2013;**20**:668–679.

difficult to identify nature of available information about medicinal uses of plants poses significant challenges to its use in pharmacology. Informatics pipelines that can integrate and link such medicinal plant knowledge with contemporary biomedical resources may provide the foundation for essential prioritization strategies that can be used to identify the most promising leads.

The slow and often expensive conventional drug discovery process for identifying plant-based medicines may be improved using computational approaches to drug design and discovery. A review of potential problems with discovery of plant-based medicines and possible computational approaches has been published recently.[8] In addition to the challenges of using conventional methods for discovery of plant-based medicines, technical problems may hinder quality control and clinical testing. For example, the identification and authentication of plant species is of primary importance. However, the ambiguity of plant names and botanical features that are used to characterize the authenticity of species in literature poses significant challenges.[12] Furthermore, botanical extracts are often mixtures of compounds, which may make the purification and identification of active ingredients difficult. Additionally, the composition of active ingredients may be affected by the weather, agronomical parameters, and processing methods. Herbal remedies may be synergistic multi-plant combinations, thus making evaluation complex. Finally, the lack of toxicity information related to medicinal herbs may also present a bottleneck in identifying potential plant-based medicines. These types of problem hinder the identification and study of potentially useful plant-based medicines.

There is a paucity of contemporary clinical information about plant-based medicines, which is a major drawback towards their standardization. To facilitate this area of research in the USA, the Center for Drug Evaluation and Research (CDER) has published guidelines that describe the unique features of botanicals and practical difficulties in their development.[13] CDER's regulatory policies were intended to encourage botanical drug development and have provided enhanced opportunities for clinical investigation.[14] Clinical information related to safety and efficacy, such as that associated with clinical trials, is important and must be accessible to the public. The success of such clinical trials may help to increase the acceptability of plant-derived drugs, and may also prove to be essential for designing future bioprospecting strategies.

The largest public repository for information about clinical trials conducted around the world is ClinicalTrials.gov.[15] ClinicalTrials.gov is a database developed by the US National Institutes of Health in collaboration with the Food and Drug Administration. It provides information on clinical trials for a wide range of diseases, conditions, and drug interventions, allowing clinicians, researchers, and patients to locate clinical trials conducted worldwide. The database contains studies that can be categorized based on conditions, drug interventions, sponsors, locations, rare diseases, and dietary supplements. As of October 2012, it contained 134 268 registered trials from 180 countries.

Plant taxonomy has been a useful guide for identifying medicinal plants and associated phytochemicals.[16] Thousands of plant species have been used traditionally for medicinal applications. Documentation of such knowledge has led to some understanding of the patterns of occurrence of medicinally important properties of plants,[17] which are a result of associated bioactive metabolites. It has been shown that closely related plant species may share similar biochemical properties,[18] and this presumption of shared biochemistry among closely related plant species has led to the field of chemosystematics.[19] Typically, certain classes of chemical metabolites are commonly found in a specific family (eg, anthraquinones in the Polygonaceae family) or in selected families from a specific order of plants.[20] Correlating the knowledge of medicinally important families with the characteristics of associated metabolites would be an important step in fine-tuning the search for drugs with specific targets, or identifying potentially new drug classes.

However, exploiting plant biodiversity for pharmacological ends may raise concerns. In particular, global decline in plant biodiversity (eg, owing to destruction of forests[21]) has damaged several medicinally important plant species.[22] The Global Convention on Biological Diversity (CBD),[23] the Convention on International Trade in Endangered Species (CITES),[24] the Forest Service, and the National Center for the Preservation of Medicinal Herbs are some of the agencies that keep track of, and protect, endangered medicinal species worldwide. Habitat destruction and overharvesting are major problems that impede conservation efforts and the sustainable use of potentially medicinal plant species.[22] Exploitation and patenting of the resources from tropical forests by private corporations also makes them unavailable to indigenous populations, thus directly affecting those communities which may have significant insights into the potential medicinal uses of local plants.[25] The initiatives of CBD and CITES have led to the promotion of conservation of medicinal plants. For example, CBD initiatives have included measures to ensure that indigenous communities receive an equitable share of benefits, to regulate the impact assessments and to assist local governments in the development of legislation to ensure that traditional knowledge is preserved.

For this case study, the following objectives were set: (1) determine the number of plant-related interventions cataloged in ClinicalTrials.gov; (2) analyze the taxonomic distribution pattern of plant species that are related to drug interventions within ClinicalTrials.gov; (3) relate drug categories from ClinicalTrials.gov with important plant families based on potential source(s) of origin of drug interventions; and (4) identify ClinicalTrials.gov plant-based interventions that may be further characterized by their conservation status. Consequently, information from a number of biodiversity resources was required to quantify the impact of medicinal plants that are associated with clinical trials (as indexed in ClinicalTrials.gov).

MATERIALS AND METHODS

The main objective of the case study was to extract and integrate plant-associated, drug-related, and clinical trial information to highlight the distribution of existing potential plant-based drug interventions and the associated drug categories used in clinical trials (as indexed in ClinicalTrials.gov). A secondary objective was to study the taxonomy of the source plants to determine any possible pattern of drug-rich plant groups. A flowchart of the workflow is depicted in figure 1.

Identification of plant-based drugs

A combined phytochemical list was developed from Dr Duke's Phytochemical and Ethnobotanical Database,[26] Kyoto Encyclopedia of Genes and Genomes (KEGG) list of phytochemicals, and the KEGG list of phytochemicals used as drugs.[27] The lists from each of these sources were merged, removing duplicates, to create a unique list of phytochemicals.

The list of 2684 drug interventions from ClinicalTrials.gov was screened through the DrugBank[28] database, which includes drug names and their synonyms. The screening resulted in two categories of clinical trial drug interventions: (1) those listed in DrugBank; and (2) those not listed in DrugBank. Drug

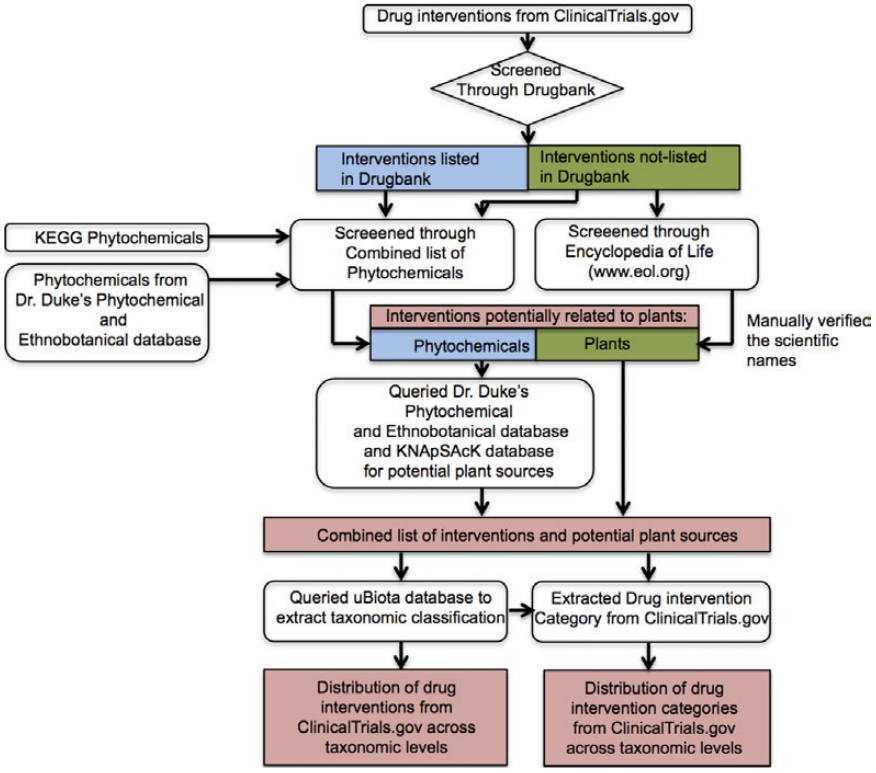

Figure 1 Workflow for identification of potential plant-derived interventions from ClinicalTrials.gov. KEGG, Kyoto Encyclopedia of Genes and Genomes.

interventions from both these categories were screened through the combined phytochemical list. This screening was used to identify phytochemical-based drug interventions in ClinicalTrials.gov. The resulting list includes only those interventions where either the name (or synonyms) matched names from the combined phytochemical list created as described above or contained plant names (either common or scientific). These phytochemical-based drug interventions were then linked to their potential plant sources by searching Dr Duke's Phytochemical and Ethnobotanical database and the KNApSAcK database (a comprehensive species–metabolite relationship database).[29] The screening of drug interventions against DrugBank and the combined phytochemical list was performed by approximate string matching as implemented in the Ruby gem 'Amatch,' using a pair-distance method. Briefly, this method considers the number of adjacent character pairs that are contained in two strings, thereby giving the advantage of accounting for the characters, and also the character ordering in the original strings. The pair-distance between two strings (s1 and s2) is calculated using the similarity metric:

$$\text{Pair distance score} = \frac{2 \times |\text{pairs}(s1) \cap \text{pairs}(s2)|}{|\text{pairs}(s1)| + |\text{pairs}(s2)|}$$

A score of 1.0 indicates an exact match. The string matching was performed at decreasing threshold values in 0.02 increments (ie, 1, 0.98, 0.96, 0.94, 0.92, 0.9) and was manually assessed at each threshold for the number of correct unique drug entities matched. It was observed that at a threshold of ≤0.94 the number of correct unique drug entities matched remained the same. Therefore, a threshold of 0.94 was chosen for this study. To identify additional potential plant-based interventions, the Encyclopedia of Life (EOL; an online catalog of life on Earth)[30] application programming interface was used within a Ruby script to identify drug interventions that contained plant scientific names. The resulting compilation of scientific names was then manually verified. The commercial availability of identified phytochemical or plant interventions was determined by searching RxNorm.[24] Furthermore, the chemical nature (eg, alkaloids, terpenes, alcohols, or flavonoids) of the identified phytochemical or plant interventions was identified by searching their respective MeSH or ChEBI hierarchy as available in PubChem.

Analyzing the taxonomic diversity of plants associated with drugs used in clinical trials

After compilation of the drug interventions and scientific names of their potential plant sources, the taxonomic distribution was analyzed. This was accomplished by extracting the taxonomy of plant species from uBiota, a locally generated unified taxonomy that is a compilation of organism taxonomy from ITIS,[31] National Center for Biotechnology Information (NCBI) Taxonomy[32] and Catalogue of Life.[33] The results were represented as a pseudo-phylogenetic tree created using FigTree V.1.3.1,[34] where the length of each branch reflected the number of potential drug interventions associated with a particular taxonomic level. The drug interventions were then categorized based on the drug categories listed in ClinicalTrials.gov and their respective trial 'phase' was also extracted. This was done via a Ruby script that leveraged a RESTful application programming interface with ClinicalTrials.gov.[7] The ClinicalTrials.gov drug interventions categories were extracted and linked with plant families. A series of Ruby scripts were then used to transform the data into Newick formatted tree files (a common file structure used for representing phylogenetic trees[35]) for the final analysis.

Assessment of extinction risk of plant species identified

The extinction risk associated with plant species identified from ClinicalTrials.gov was assessed by mapping the species against the International Union for the Conservation of Nature (IUCN) Red List of Threatened Species version 2012.1 (the 'Red

List').[36] A search was conducted on IUCN that included all the species under the taxonomic category 'Plantae' and assessment categories extinct (EX), extinct in the wild (EW), critically endangered (CR), endangered (EN), vulnerable (VU), lower risk/conservation-dependent (LR/cd), near threatened (NT), or lower risk/near threatened (LR/nt) (see figure 2 for a taxonomy of major assessment categories). Species names of medicinal plants identified from ClinicalTrials.gov drug interventions were searched for within an export of the Red List species names and associated taxonomic classification.

RESULTS

Status of plant-based interventions in ClinicalTrials.gov

The main objective of the case study was to identify potential plant-related interventions for which clinical trials are registered in the ClinicalTrials.gov database (as of February 2012). A summary of results is presented in table 1. From the list of 2684 drug interventions in ClinicalTrials.gov, 1314 (49%) could be mapped to DrugBank. The list of drug interventions from ClinicalTrials.gov and their synonyms were then used to screen the list of phytochemicals compiled from Dr Duke's Phytochemical and Ethnobotanical database and KEGG databases (phytochemicals and list of phytochemicals used as drugs). From all the drug interventions in ClinicalTrials.gov, 293 (~11%) could be associated with a phytochemical based on this compiled list. The result from this analysis includes only those interventions that could be mapped to a phytochemical entity via names (scientific or common) or synonyms. An additional 114 drug interventions could be associated with plants based on screening through EOL (~4% of all drug interventions in ClinicalTrials.gov). In total, 407 interventions in ClinicalTrials.gov were potentially related to plants (ie, ~15% of all drug interventions in ClinicalTrials.gov).

Distribution of plant-based interventions across different taxonomic levels

The second objective of this case study was to analyze the taxonomic distribution of origin sources for potential plant-related drug interventions from ClinicalTrials.gov. For the 407 drug interventions identified as plant-based, 1226 plant species were located that might potentially serve as a source of origin (about three plant species for each drug intervention). Several plant species could potentially be associated with more than one drug intervention. These plant species were then matched with their respective taxonomic classification listed in uBiota. The complete list of plant species associated with drug interventions together with their taxonomic classification is provided in online supplementary table S1. The 1226 plant species located were distributed across nine of the 20 divisions listed in uBiota. The division Magnoliophyta alone was associated with 398 of the 407 plant-based drug interventions. The species of origin of drug interventions were distributed among 10 of 41 classes, with Magnoliopsida being the major class. Seventy-two of 251 orders represent the 1226 plant species. These plant species were distributed across 175 families of the 941 families in kingdom Plantae. The top 10 plant families with an association with drug interventions in clinical trials were: Fabaceae, Asteraceae, Apiaceae, Rosaceae, Solanaceae, Poaceae, Lamiaceae, Rutaceae, Brassicaceae, and Amaryllidaceae. These top 10 families accounted for 246 out of the total 407 plant-based interventions identified in this study (~60%). Figure 3 shows the combined distribution of drug-producing plant species in different families, orders, and classes. The branch length in the figure is directly proportional to the number of potentially associated drugs. The top 10 genera that were identified as potential sources of origin for drug interventions were (families shown in parentheses): *Glycine* (Fabaceae); *Citrus* (Rutaceae); *Allium* (Amaryllidaceae); *Prunus* (Rosaceae); *Zea* (Poaceae); *Apium* (Apiaceae); *Panax* (Araliaceae); *Theobroma* (Sterculiaceae); *Solanum* (Solanaceae); *Camellia* (Theaceae); and *Urtica* (Urticaceae) (table 2).

Table 1 Status of plant-based interventions in ClinicalTrials.gov

	Plant/total	Percentage
Phytochemical associated	293/2684	11
Plants	114/2684	4
Total plant-based interventions listed	407/2684	15
Plant-based interventions in phase 0	1/407	0.2
Plant-based interventions in phase 1	21/407	5
Plant-based interventions in phase 2	70/407	17
Plant-based interventions in phase 3	74/407	18
Plant-based interventions in phase 4	218/407	54
Phase information not provided	23/407	6

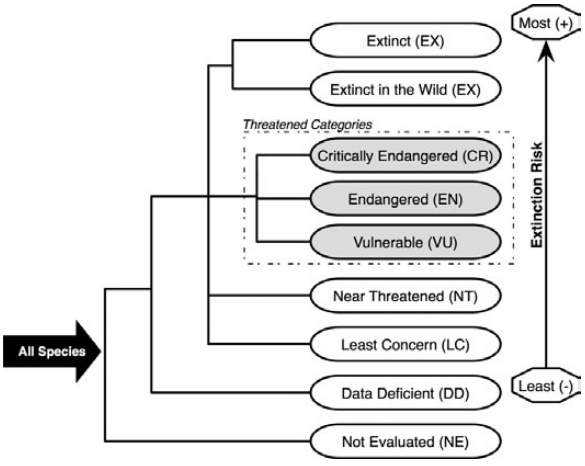

Figure 2 International Union for the Conservation of Nature Red List categories (Source: http://www.iucnredlist.org).

Distribution of plant-based interventions across different drug-categories

The third objective of this case study was to determine the distribution of plant-related drug interventions across different drug categories listed in ClinicalTrials.gov. Forty-five drug intervention categories were listed in ClinicalTrials.gov and most interventions were associated with more than one category. The plant-based interventions identified in this study were represented in all 45 categories, although the number might have been small for some of the categories (eg, two for antisickling agents and four for natriuretic agents). The top three major categories based on the numbers of associated drug interventions were (1) anti-infective agents (236 drug interventions); (2) antineoplastic agents (225 drug interventions); and (3) micronutrients (200 drug interventions). A comparison of distribution of plant-related drug interventions and non-plant based drug interventions across different drug categories is shown in figure 4. The important plant families of potential sources of origin for drug interventions included within different categories are listed in table 3.

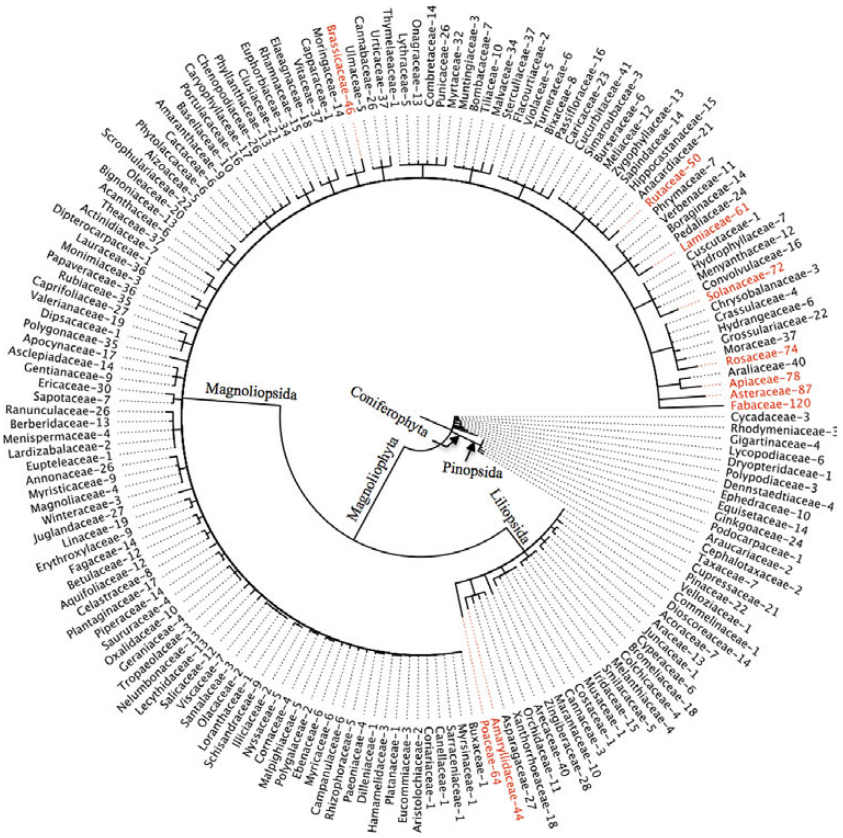

Figure 3 Distribution of plant-based drug interventions from ClinicalTrials.gov across different taxonomic levels (branch length is directly proportional to the number of potentially associated drug interventions). The top 10 families are highlighted.

Assessment of extinction risk of medicinally important plant species identified

The plant species identified as being potentially related to at least one of the interventions listed in ClinicalTrials.gov were mapped to the IUCN Red List of threatened species version 2012.1 for assessment of conservation status. Of all the plant species, 45 (~3.7%) were identified as falling into one of the categories of Red List conservation threat (table 4). Among the 45 identified plant species, 80% fall into these three major categories of extinction risk (CR, EN, or VU, as shown in figure 2). The percentage contribution of plant species for all the categories is shown as a pie chart in figure 5.

DISCUSSION

The linking of heterogeneous data sources to identify pharmaceutical leads from existing knowledge bases poses significant opportunities for biomedical data mining. However, there are several challenges in linking traditionally unlinked biomedical and biodiversity data sources. Taxonomic (scientific) names have been used as key identifiers for integrating information between biomedical and biodiversity domains.[1] Previous studies that have attempted to link genetic and biodiversity information have shown promise for studying the evolution and spread of infectious diseases.[37 38] There may be the potential to leverage chemical diversity in nature for drug discovery. To this end, it might be useful to mine drug-related information (chemical nature, mechanism of action, clinical efficacy) in conjunction with the species diversity information to identify potential sources for chemical ingredients. Developing bridges between biodiversity and biomedical knowledge can enable synergistic advances—for example, (1) using knowledge of shared chemistry in drug exploration; and (2) identifying and locating species with potential, and prioritizing conservation strategies accordingly. This study attempted to explore the feasibility of using chemical names in conjunction with names of species to link information from different sources. The discussion will thus focus on the perspectives gained from the medicinal plant case study presented.

Table 2 Taxonomic distributions of plant species associated with interventions listed in ClinicalTrials.gov

Taxonomic level	Count
Division	9
Class	10
Order	72
Family	175
Genus	734
Species	1226

Information about plant-based therapies is generally difficult to identify or scattered and embedded within text. This knowledge may be embedded within a non-specific category. As a part of the case study, a goal was set to explore ClinicalTrials.gov for potential plant-related interventions and understand the taxonomic distribution of associated species. The primary challenge faced was the difficulty of linking drugs to their respective plant species of origin. The plant-based interventions listed in ClinicalTrials.gov are either generic names of drugs or common names of plants. There are numerous challenges in extracting plant names from existing data sources, and the plant names may vary across resources, especially the common (vernacular) names listed. Such variation, further complicated by frequent misspellings or typographical errors, may result in incorrect identification of plant materials. To deal with such challenges, 'taxonomically intelligent' strategies are required.[1 2]

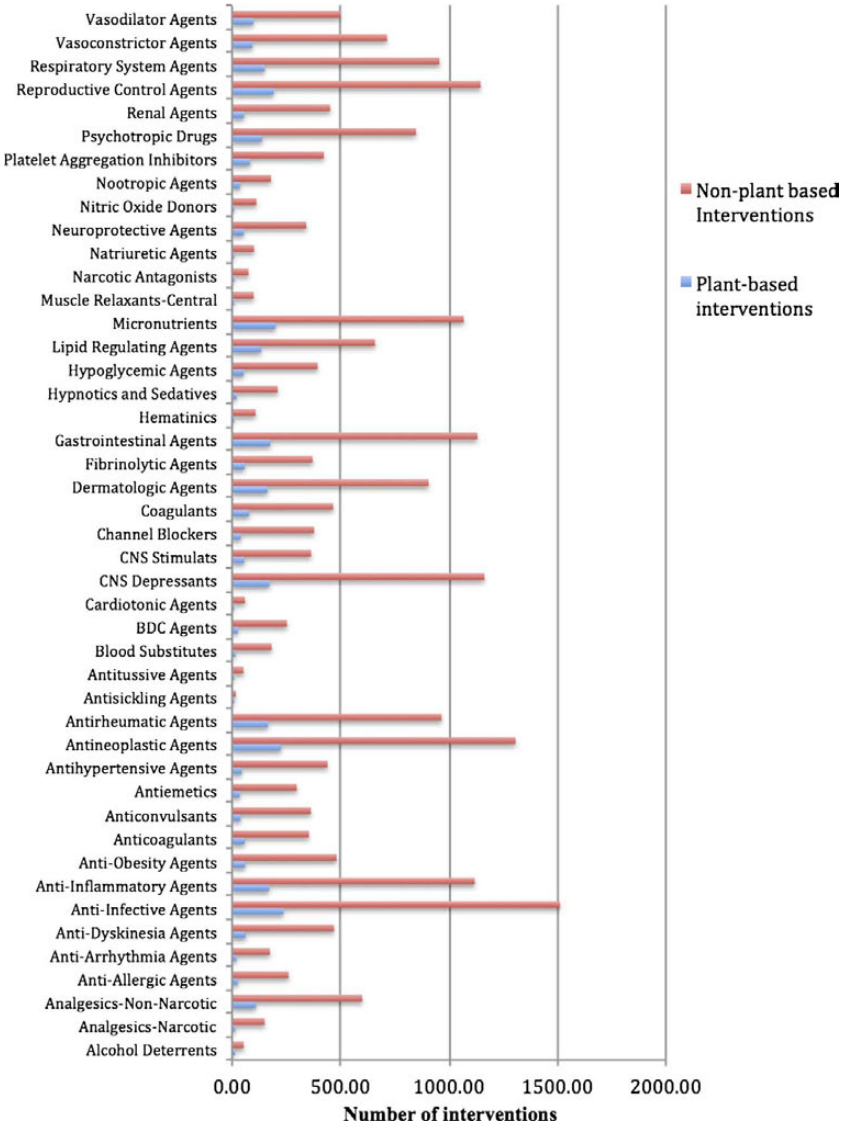

Figure 4 Distribution of drug interventions across different drug categories in ClinicalTrials.gov.

Within ClinicalTrials.gov, the lack of taxonomically intelligent indexing may lead to misleading identification of potential medicinal plants. For example, in searching for clinical trials in ClinicalTrials.gov indexed by '*Rhamnus frangula*' (which is the alder buckthorn plant) results are returned for *Hippophae* sp (another plant called sea buckthorn): NCT00767156, NCT01697085, and NCT00739713. It is important to deal with such problems while indexing medicinal plant information. Incorrect or misleading identification of plant species may reflect ineffective treatment and may potentially lead to adverse reactions (including fatality). There are documented instances where such nomenclature confusion has led to serious consequences. For example, a case of renal failure was reported in Europe that was due to poisoning as a result of incorrect use of *Aristolochia fangchi*[39] or neurotoxicity resulting from *Illicium anisatum* in a herbal tea mixture.[40]

Additional challenges in taxonomic nomenclature make the indexing of medicinal plant information even more difficult. Of particular concern is the use of alternative names ('synonyms') for the same species. For example, *Rhamnus frangula* is a synonym for the accepted name for the species *Frangula dodonei*. There are several reasons that can lead to synonyms resulting in conflicting taxonomic categorization (as detailed by Fenneman[41]). In an effort to deal with this concern, the Royal Botanic Gardens, Kew is pioneering an initiative (the Medicinal Plant Names Index (MPNI)) to connect medicinal plant common names, accepted scientific name(s), and synonyms—with the ultimate goal of creating an interlinked, user-friendly, comprehensive map of medicinal plants.

This study highlights the importance of developing resources such as MPNI for future bioprospecting studies. In particular, this study showed that some of the resources providing plant species names for identified clinical trial intervention do not have a comprehensive list of synonyms or adequate indexing by accepted scientific names. As a result, some of the names used in this study (listed in the online supplementary tables) may be synonyms and not the currently accepted names for particular species. With the availability of EOL, obtaining the standard taxonomic names was possible. However, the absence of chemical names of drugs or ingredients and identification of plant origin required additional resources—namely, (1) DrugBank for chemical information related to drugs; and, (2) KEGG, Dr Duke's Phytochemical and Ethnobotanical Database and KNApSAcK for determination of whether a chemical was potentially of plant origin. However, robust linking of resources to identify potential phytochemical therapeutic agents would

Table 3 Top five plant families associated with different drug intervention categories

Drug category	Commercially available	Under consideration	Potential plant families
Alcohol deterrents	10	1	Fabaceae, Solanaceae, Rosaceae, Sterculiaceae, Rubiaceae
Analgesics—narcotic	12	2	Fabaceae, Solanaceae, Papaveraceae, Rosaceae, Caricaceae
Analgesics—non-narcotic	79	28	Fabaceae, Apiaceae, Asteraceae, Rosaceae, Solanaceae
Anti-allergic agents	28	5	Fabaceae, Apiaceae, Cucurbitaceae, Poaceae, Asteraceae
Anti-arrhythmia agents	18	2	Fabaceae, Rosaceae, Rubiaceae, Apiaceae, Arecaceae
Anti-dyskinesia agents	46	12	Fabaceae, Rosaceae, Asteraceae, Apiaceae, Solanaceae
Anti-infective agents	167	55	Fabaceae, Asteraceae, Apiaceae, Rosaceae, Solanaceae
Anti-inflammatory agents	127	40	Fabaceae, Apiaceae, Asteraceae, Rosaceae, Solanaceae
Anti-obesity agents	44	7	Fabaceae, Asteraceae, Rosaceae, Apiaceae, Papaveraceae
Anticoagulants	45	7	Fabaceae, Rosaceae, Apiaceae, Asteraceae, Poaceae
Anticonvulsants	34	5	Fabaceae, Rosaceae, Apiaceae, Asteraceae, Araliaceae
Antiemetics	30	6	Fabaceae, Rosaceae, Apiaceae, Asteraceae, Solanaceae
Antihypertensive agents	34	4	Fabaceae, Rosaceae, Poaceae, Arecaceae, Solanaceae
Antineoplastic agents	156	58	Fabaceae, Asteraceae, Apiaceae, Rosaceae, Poaceae
Antirheumatic agents	123	39	Fabaceae, Apiaceae, Asteraceae, Rosaceae, Solanaceae
Antisickling agents	2	0	Apocynaceae, Taxaceae
Antitussive agents	4	0	Ephedraceae, Sterculiaceae, Malvaceae, Ranunculaceae, Rhamnaceae
Blood substitutes	17	0	Rosaceae, Punicaceae, Araliaceae, Fabaceae, Malvaceae
Bone density conservation agents	20	2	Fabaceae, Apiaceae, Asteraceae, Rosaceae, Poaceae
Cardiotonic agents	7	1	Rosaceae, Fabaceae, Annonaceae, Sterculiaceae, Araliaceae
Central nervous system depressants	124	37	Fabaceae, Rosaceae, Asteraceae, Apiaceae, Solanaceae
Central nervous system stimulants	49	8	Fabaceae, Rosaceae, Poaceae, Araliaceae, Asteraceae
Channel blockers	33	6	Fabaceae, Apiaceae, Poaceae, Rosaceae, Asteraceae
Coagulants	63	12	Fabaceae, Apiaceae, Rosaceae, Asteraceae, Solanaceae
Dermatologic agents	129	25	Fabaceae, Asteraceae, Apiaceae, Rosaceae, Poaceae
Fibrinolytic agents	45	6	Fabaceae, Apiaceae, Poaceae, Rosaceae, Asteraceae
Gastrointestinal agents	133	33	Fabaceae, Apiaceae, Rosaceae, Solanaceae, Asteraceae
Hematinics	8	0	Araliaceae, Cucurbitaceae, Punicaceae, Fabaceae, Rosaceae
Hypnotics and sedatives	18	1	Fabaceae, Solanaceae, Rosaceae, Asteraceae, Apiaceae
Hypoglycemic agents	43	12	Fabaceae, Rosaceae, Asteraceae, Apiaceae, Amaryllidaceae
Lipid regulating agents	99	32	Fabaceae, Apiaceae, Asteraceae, Rosaceae, Poaceae
Micronutrients	144	52	Fabaceae, Asteraceae, Apiaceae, Rosaceae, Solanaceae
Muscle relaxants—central	7	0	Ephedraceae, Papaveraceae, Aquifoliaceae, Plantaginaceae, Theaceae
Narcotic antagonists	10	1	Fabaceae, Araliaceae, Rosaceae, Cannabaceae, Punicaceae
Natriuretic agents	4	0	Cucurbitaceae, Apocynaceae, Apiaceae, Asteraceae, Fabaceae
Neuroprotective agents	46	12	Fabaceae, Asteraceae, Rosaceae, Solanaceae, Poaceae
Nitric oxide donors	5	2	Fabaceae, Rosaceae, Asteraceae, Punicaceae, Cucurbitaceae
Nootropic agents	29	7	Fabaceae, Rosaceae, Solanaceae, Apiaceae, Asteraceae
Platelet aggregation inhibitors	59	21	Fabaceae, Rosaceae, Asteraceae, Apiaceae, Solanaceae
Psychotropic drugs	108	24	Fabaceae, Rosaceae, Asteraceae, Apiaceae, Solanaceae

Continued

Table 3 Continued

Drug category	Commercially available	Under consideration	Potential plant families
Renal agents	44	5	Fabaceae, Apiaceae, Asteraceae, Solanaceae, Rosaceae
Reproductive control agents	150	39	Fabaceae, Apiaceae, Asteraceae, Rosaceae, Solanaceae
Respiratory system agents	119	26	Fabaceae, Asteraceae, Solanaceae, Apiaceae, Rosaceae
Vasoconstrictor agents	78	12	Fabaceae, Asteraceae, Rosaceae, Apiaceae, Solanaceae
Vasodilator agents	83	11	Fabaceae, Rosaceae, Solanaceae, Asteraceae, Apiaceae

require the use of uniform chemical identifiers (which are missing from many of the resources used in the case study) in addition to standardized taxonomic names. The use of standard International Union of Pure and Applied Chemistry International Chemical Identifier (InChI) to encode chemical substances can facilitate the search and retrieval of chemical information in databases.[42] Standard chemical databases, such as ChEBI and PubChem, provide InChIs for their contents. Additionally, DrugBank provides the InChI key for many of the chemical ingredients that it indexes. However, the absence of such unified identifiers or normalized names in other databases storing species-metabolite information (eg, KNApSAcK) poses additional challenges in cross-domain data integration.

The approach used in this study thus leveraged approximate string matching for screening of chemicals using their names or synonyms. However, a limitation of this approach may be that variations of chemical names may not be contained within the list of synonyms. Future work may leverage more advanced named entity recognition tools for identifying chemical names (eg, OSCAR4[43]). Another limitation of this study was that interventions were only identified when their names or synonyms could be matched to the combined phytochemical list. Possibly, therefore, drug interventions that were plant-derived semi-synthetic or plant-product mimic synthetic were missed if their source ingredient or precursor was not mentioned as a synonym. Plant secondary metabolites provide valuable precursors that may be pharmacologically important. For example, some of the interventions in ClinicalTrials.gov derived from plant metabolites after chemical modifications are (precursor plant(s) in parenthesis): acitretin (*Daucus carota*), bimatoprost (*Allium sativum*, *Artemisia dracunculus*), Coarsucam (*Artemisia annua*, *Cinchona officinalis*), docetaxel (*Taxus wallichiana*, *Taxus baccata*, *Taxus brevifolia*), and nitisinone (*Callistemon citrinus*). In addition to the semisynthetic drugs, plant secondary metabolites can also provide guiding molecules for development of synthetic drugs and mimics—for example, Abraxane (*Taxus wallichiana*, *Taxus brevifolia*), betaxolol HCl (*Ephedra sinica*, *Acacia rigidula*), gefitinib (*Zea mays*), and lapatinib (*Zea mays*, *Cocos nucifera*).

Results from the assessment of the results for the case study suggest that ~15% of drug interventions listed in ClinicalTrials.gov are plant-related. Many of these botanical interventions are commercially available but not as 'drugs'. They are available as dietary supplements without specific disease treatment claims. Such nutraceutical interventions do not require FDA approval. This highlights the relatively low volume of clinical studies that focus on plant-based interventions, and thus implies a significant opportunity for development of phyto-therapies. However, it is important to note that complete characterization of phytochemical components in botanical-based interventions is challenging, especially when searching for those that may be of medicinal value. In particular, it can be complicated because phytochemical properties can vary from batch to batch, depending on plant growing conditions and plant part(s) used. This type of problem significantly affects the potential viability of phytochemical agents for human trials. Similarly, some traditional herbal remedies are mixtures of extracts from several plants. Assessment of the medicinal viability of such multi-plant combinations requires factorial trials that become more complex as the number of potential ingredients increases. Finally, the lack of standardized authentication and toxicological evidence adds a further barrier to the clinical assessment of plant-based remedies. Chen *et al*[14] have reviewed these topics and related regulatory policies.

Table 4 List of plant species that are at risk of extinction

Species Id	Class	Order	Family	Species	Status
32986	Magnoliopsida	Lecythidales	Lecythidaceae	*Bertholletia excelsa*	VU
30803	Magnoliopsida	Theales	Dipterocarpaceae	*Cotylelobium scabriusculum*	CR
34171	Magnoliopsida	Asterales	Compositae	*Dendroseris neriifolia*	CR
33203	Magnoliopsida	Ebenales	Ebenaceae	*Diospyros celebica*	VU
33048	Magnoliopsida	Ebenales	Ebenaceae	*Diospyros crassiflora*	EN
30804	Magnoliopsida	Theales	Dipterocarpaceae	*Dipterocarpus glandulosus*	CR
30805	Magnoliopsida	Theales	Dipterocarpaceae	*Dipterocarpus hispidus*	CR
30806	Magnoliopsida	Theales	Dipterocarpaceae	*Dipterocarpus insignis*	CR
30807	Magnoliopsida	Theales	Dipterocarpaceae	*Dipterocarpus zeylanicus*	EN
31280	Magnoliopsida	Eucommiales	Eucommiaceae	*Eucommia ulmoides*	LR/nt
38148	Magnoliopsida	Sapindales	Rutaceae	*Flindersia laevicarpa*	VU
162168	Liliopsida	Liliales	Amaryllidaceae	*Galanthus nivalis*	NT
46671	Liliopsida	Orchidales	Orchidaceae	*Gastrodia elata*	VU
32353	Ginkgoopsida	Ginkgoales	Ginkgoaceae	*Ginkgo biloba*	EN
33701	Magnoliopsida	Sapindales	Zygophyllaceae	*Guaiacum officinale*	EN
32955	Magnoliopsida	Sapindales	Zygophyllaceae	*Guaiacum sanctum*	EN
30887	Magnoliopsida	Theales	Dipterocarpaceae	*Hopea brevipetiolaris*	CR
30808	Magnoliopsida	Theales	Dipterocarpaceae	*Hopea cordifolia*	EN
32982	Magnoliopsida	Celastrales	Aquifoliaceae	*Ilex paraguariensis*	LR/nt
63495	Magnoliopsida	Juglandales	Juglandaceae	*Juglans regia*	NT
38016	Magnoliopsida	Ebenales	Sapotaceae	*Madhuca microphylla*	EN
34963	Magnoliopsida	Magnoliales	Magnoliaceae	*Magnolia officinalis*	LR/nt
33537	Magnoliopsida	Ebenales	Sapotaceae	*Palaquium grande*	VU
34402	Magnoliopsida	Laurales	Lauraceae	*Persea schiedeana*	VU
43981	Liliopsida	Arecales	Palmae	*Phytelephas aequatorialis*	NT
38910	Magnoliopsida	Sapindales	Simaroubaceae	*Picrasma excelsa*	VU
34189	Coniferopsida	Coniferales	Pinaceae	*Pinus gerardiana*	LR/nt
39068	Coniferopsida	Coniferales	Pinaceae	*Pinus palustris*	VU
63497	Magnoliopsida	Sapindales	Anacardiaceae	*Pistacia vera*	NT
33631	Magnoliopsida	Rosales	Rosaceae	*Prunus africana*	VU
33190	Magnoliopsida	Fabales	Leguminosae	*Pterocarpus angolensis*	LR/nt
34620	Magnoliopsida	Fabales	Leguminosae	*Pterocarpus marsupium*	VU
63485	Magnoliopsida	Sapindales	Anacardiaceae	*Rhus coriaria*	VU
31852	Magnoliopsida	Santalales	Santalaceae	*Santalum album*	VU
30817	Magnoliopsida	Theales	Dipterocarpaceae	*Shorea affinis*	EN
30818	Magnoliopsida	Theales	Dipterocarpaceae	*Shorea congestiflora*	CR
30823	Magnoliopsida	Theales	Dipterocarpaceae	*Shorea ovalifolia*	CR
30824	Magnoliopsida	Theales	Dipterocarpaceae	*Shorea trapezifolia*	CR
30826	Magnoliopsida	Theales	Dipterocarpaceae	*Shorea zeylanica*	CR
30889	Magnoliopsida	Theales	Dipterocarpaceae	*Stemonoporus canaliculatus*	CR
30836	Magnoliopsida	Theales	Dipterocarpaceae	*Stemonoporus reticulatus*	EN
33062	Magnoliopsida	Myrtales	Combretaceae	*Terminalia ivorensis*	VU
30840	Magnoliopsida	Theales	Dipterocarpaceae	*Vatica affinis*	CR
33959	Magnoliopsida	Magnoliales	Myristicaceae	*Virola surinamensis*	EN
37083	Magnoliopsida	Ebenales	Sapotaceae	*Vitellaria paradoxa*	VU

CR, critically endangered; EN, endangered; LR/NT, lower risk/near threatened; NT, near threatened; VU, vulnerable.

Examination of the chemical nature of the compounds identified in this study showed that the major categories were alkaloids (eg, ajmaline, physostigmine), flavonoids (eg, epigallocatechin, hesperidin), coumarins (eg, ficusin, methoxsalen), steroids (eg, digoxin, phytosterol), terpenoids (eg, stevioside, parthenolide), amino acids and derivatives (eg, glycine, creatine). Of these, the largest group was alkaloids, which are associated with a diverse range of drugs, including those that are stimulant (eg, caffeine, nicotine), anti-bacterial (eg, berberine), anti-hypertensives (eg, reserpine), and anti-cancer (eg, vincristine). A list of the chemical nature of the putative phytochemical interventions identified in this study is included in online supplementary table S7. The chemical categories and names listed may not completely reflect the medicinal plant potential because this case study used ClincialTrials.gov as its primary source of biomedical knowledge of medicinal plant use. However, use of ClinicalTrials.gov was relevant as we were interested in identifying plant-based treatments that might have been accepted and validated or might have received considerable attention towards validation in light of contemporary scientific methodologies.

Mapping drug interventions from ClinicalTrials.gov to the taxonomy of potential plant sources of origin helped to disclose potentially significant medicinally important plant families.

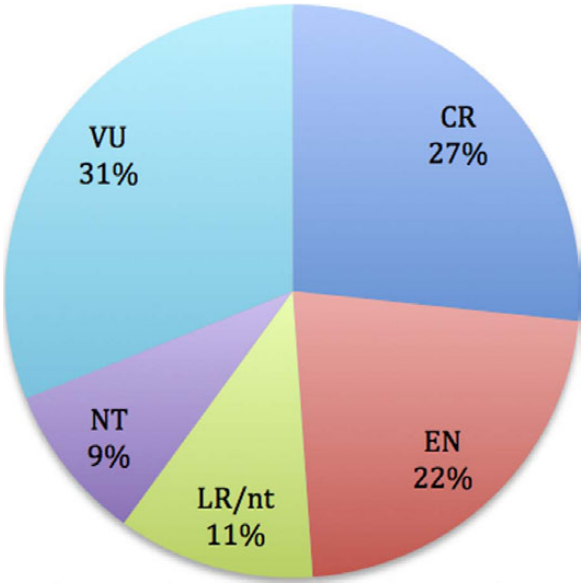

Figure 5 Percentage contribution of plant species among different categories listed in the International Union for the Conservation of Nature Red List. CR, critically endangered; EN, endangered; LR/nt, lower risk/near threatened; NT, near threatened; VU, vulnerable.

To provide a comprehensive taxonomical resource, this study used the uBiota taxonomy that unifies taxonomic information from the Integrated Taxonomic Information System (ITIS), NCBI Taxonomy, and the Catalogue of Life. Based on the results of the case study, Fabaceae was the plant family associated with the greatest number of drug interventions. Fabaceae is the second largest family of medicinal plants and has been described in the Chinese and Japanese Pharmacopoeia.[44] It is important to note that ~60% of identified plant-based drug interventions are clustered within the top 10 families. The findings are plausible since Fabaceae, Solanaceae, Poaceae, Asteraceae, and Amaryllidaceae have been shown to be sources of important drugs.[23] However, a difficulty in interpreting this result is that it is unclear whether the distribution pattern is biased because of limited number of plant-based interventional studies in ClinicalTrials.gov. Therefore, to better quantify the effect of plants in contemporary medicine, it will be necessary to carry out a more comprehensive analysis of available drugs (eg, as cataloged in resources such as RxNorm[24]) and chemical product databases (eg, the Combined Chemical Dictionary Database online[45]).

Although several potential medicinal plant species were identified in this study, drug discovery remains difficult. A significant challenge is that there are limited standardized data that can be used for accurate identification and authentication of plant species. The use of standardized 'DNA barcodes' (which are community-supported molecular signatures for species identification, for plants: a combination of information from the rbcL and matK genes) does promise to be an efficient and reliable method for authenticating plant species.[46] In support of this, the Medicinal Materials DNA Barcode Database has been developed for molecular sequence-based querying of medicinal plant information and provides some bioinformatics tools for searching and sequence comparison.[47] However, at least at the time writing, the DNA barcode data specific for medicinal plants are limited.

Forty-five drug intervention categories within ClinicalTrials.gov are used to describe the nature of interventions. The case study showed that plant-based drug interventions are represented in all 45 categories (figure 4). To some extent, this implicates the potential of plants to serve as sources for a wide variety of drugs if due attention were to be invested in exploring this area of research. The interventions included within the drug categories as indexed in ClinicalTrials.gov are listed in online supplementary tables S2 (plant-based) and S3 (non-plant based). The clinical trial phase of investigations associated with plant-based drug interventions is provided in online supplementary table S5. Anti-infective agents are the top ClinicalTrials.gov category, having the most plant-based drug interventions that are in phase 3 or phase 4 clinical trials. Anti-infective agents, antineoplastic agents, and micronutrients are the top three major ClinicalTrials.gov categories with potential plant-based drug interventions. Plants have immense ability to synthesize anti-infective compounds and have been historically used effectively in traditional medicine. This is in contrast to the composition of anti-microbial agents commonly used in contemporary medicine, which are derived from bacteria and fungi.[48] Indeed, plant-based drug interventions only comprised 13.5% of total drug interventions from the anti-infective agents category.

Some of the major groups of anti-infective chemicals from plants are alkaloids (diterpenoid alkaloids, commonly isolated from the plants of the Ranunculaceae (the buttercup family) are commonly found to have antimicrobial properties), lectins, terpenoids, coumarins, tannins, flavones, and quinones. Although phenols indicate a broader category overlapping with the above-mentioned chemical groups of phytochemicals, some of the anti-infective agents comprise single substituted phenolic rings. Cinnamic and caffeic acids are common representatives of a wide group of phenylpropane-derived compounds. Catechol and pyrogallol are hydroxylated phenols. The anti-infective properties of plants have been reviewed in detail by Cowan.[48]

Antineoplastic activity in plants has also been acknowledged. For example, Hartwell compiled a list of more than 3000 plants that have been reported to be used for cancer treatment in traditional plant-based medicine.[49] Plant-based compounds have been an important source of several clinically useful anti-cancer agents such as vinblastine, vincristine, camptothecin derivatives, topotecan and irinotecan, etoposide, epipodophyllotoxin, and paclitaxel.[50] In recent years, cancer has received more attention in the research into developing plant-based interventions.[14] In this study, elemental forms of micronutrients (eg, copper or zinc) were not included as potentially plant-based. Nonetheless, micronutrients interestingly turned out to be the third major ClinicalTrials.gov category with potential plant-based interventions. It is widely accepted that consumption of fruits and vegetables is good for health as they are rich sources of vitamins and minerals. Furthermore, the use of dietary supplements is also common (eg, about 40% of the US population use multivitamins or multiminerals, which are often botanically based[51]). The inclusion of micronutrients along with other drug categories in ClinicalTrials.gov is often to test their efficacy as supplements with other primary interventions, their preventive effects, and, to some extent, their interactions with the major intervention.

The indexing of drug interventions within a given category as extracted from ClinicalTrials.gov has some limitations. The categorization is based on different studies related to a particular intervention. When a given study uses more than one intervention, it can lead to false inclusion of different interventions within the same category, even though they are not the primary intervention used for a specific condition. Future work may require the use of natural language processing systems to extract the primary interventions and their therapeutic use(s) from text associated with studies registered in ClinicalTrials.gov.

The medicinal plant case study presented here provides insight into the taxonomic diversity of potential plant-based interventions and underscores the importance of conservation efforts for plants with potentially significant medicinal value. For example, the intervention 'nordihydroguaiaretic acid,' a phenolic lignan listed in clinical trials is associated with eight studies (NCT00678015, NCT00313534, NCT00404248, NCT00057512, NCT00664677, NCT00664586, NCT00154089, NCT00259818) and is in early (phase 1 or phase 2) clinical trials for a variety of conditions: prostate cancer, brain and central nervous system tumors, myeloid and lymphocytic leukemia, cervical neoplasia, and cancer. When linked with the species-metabolite database, two source plant species were identified for nordihydroguaiaretic acid: *Guaiacum officinale and Guaiacum sanctum*. These two plant species belong to the family Zygophyllaceae. *Guaiacum officinale* is also a source for other potential medicinal phytochemical agents such as triterpenoid saponin, guaianin from the flowers,[52] as well as two saponins (guaiacin A and B) from the leaves.[53] Although this plant species has potential medicinal importance as a source for several prospective drugs, it is endangered. Highlighting the medicinal importance and assessment of conservation status might strengthen efforts for designing conservation and sustainable use strategies.

The phytochemical profile of plants is complex, making it difficult to fully understand the mechanisms of action for plant-based remedies. Chemical fingerprinting and bioactive metabolite determination may provide some insight into the chemical and biological activity of plant-based medicines.[54] Characterization of chemical constituents and their bioactivities can be important for standardizing herbal therapeutic approaches in light of potential clinical impact. This aspect can also be important for quality control. However, the isolation and identification of bioactive constituents is challenging, making the clinical assessment of plant-based drugs difficult. To be successfully used, plant-based drugs and remedies must be shown to be safe and effective. The difficulties in characterizing the chemical profiles and translating traditional knowledge into a testable hypothesis are two of the challenges faced in the development of clinically accepted plant-based drugs.[14] Thus, amidst the great potential to identify possible phyto-therapies by combining biodiversity and biomedical knowledge sources, it is important to underscore that the challenges in phyto-therapy validation should not be underestimated.

In addition to access to taxonomic information, using the organism name as shared identifiers, an array of information could be gathered (ie, genetic, geographic, morphological, etc) that might be used to uncover patterns at the molecule and species level to test comparative biology hypotheses.[1] Future work will thus focus on evaluating the potential to incorporate the array of available data that might be useful for bioprospecting applications (eg, to help identify potential phyto-therapeutic rich regions that may cluster according to disease type). Furthermore, such additional information, together with an assessment of the possible utility of medicinal plant species, might help in prioritizing and designing conservation plans. As an example, the study identified medicinally important plant species that could be mapped to the IUCN Red List. Thus, although the case study does demonstrate the ability to link biomedical and biodiversity knowledge resources, it provides only a cursory view of the potential to leverage the vast amount of information available about medicinal plants.

CONCLUSION

The goal of this study was to evaluate the feasibility of linking biodiversity and biomedical resources to enable the identification of potential phyto-therapies. Such cross-domain pipelines will be important for exploring the array of nature-derived drug sources. Through a case study, the impact of medicinal plants on drug interventions indexed in ClinicalTrials.gov was assessed. The results suggest that there is a paucity (~15%) of clinical trial studies that are associated with plant-based medicines. Nonetheless, the results reveal some medicinally important plant families and genera associated with drug interventions in ClinicalTrials.gov. This study describes an approach for identifying potentially useful taxonomic information about plant species that might be important with respect to their potential as sources of drugs. Such information coupled with structural and physicochemical properties of bioactive plant metabolites may enable a more targeted and efficient bioprospecting strategy.

Contributors VS and INS designed the study, analyzed the results, and drafted the manuscript together.

Funding National Institutes of Health. This study used resources that were funded in part by DHHS/NIH/NLM R01LM009725.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The detailed results are provided as supplementary tables (an Excel file) that will be provided to *JAMIA*.

REFERENCES

1 Sarkar IN. Biodiversity informatics: organizing and linking information across the spectrum of life. *Brief Bioinform* [Review] 2007;8:347–57.
2 Sarkar IN. Biodiversity informatics: the emergence of a field. *BMC Bioinform* [Introductory Research Support, Non-U.S. Gov't] 2009;10(Suppl 14):S1.
3 Soberon J, Peterson AT. Biodiversity informatics: managing and applying primary biodiversity data. *Philos Trans R Soc Lond B Biol Sci* [Review] 2004;359:689–98.
4 Chen ES, Sarkar IN. Towards structuring unstructured GenBank metadata for enhancing comparative biological studies. *AMIA Summits Transl Sci Proc* 2011;2011:6–10.
5 Sarkar IN. Leveraging biomedical ontologies and annotation services to organize microbiome data from Mammalian hosts. *AMIA Annu Symp Proc* 2010;2010:717–21.
6 De Luca V, Salim V, *et al*. Mining the biodiversity of plants: a revolution in the making. *Science* [Research Support, Non-U.S. Gov't] 2012;336:1658–61.
7 Buenz EJ, Schnepple DJ, Bauer BA, *et al*. Techniques: bioprospecting historical herbal texts by hunting for new leads in old tomes. *Trends Pharmacol Sci* [Historical Article Research Support, Non-U.S. Gov't Review] 2004;25:494–8.
8 Sharma V, Sarkar IN. Bioinformatics opportunities for identification and study of medicinal plants. *Brief Bioinform*. Published Online First: 15 May 2012. doi:10.1093/bib/bbs021
9 Kang TH, Moon E, Hong BN, *et al*. Diosgenin from Dioscorea nipponica ameliorates diabetic neuropathy by inducing nerve growth factor. *Biol Pharm Bull* 2011;34:1493–8.
10 Miner J, Hoffhines A. The discovery of aspirin's antithrombotic effects. *Tex Heart Inst J* [Biography Historical Article Review] 2007;34:179–86.
11 Li JW, Vederas JC. [Drug discovery and natural products: end of era or an endless frontier?]. *Biomed Khim* [Review] 2011;57:148–60.
12 Zhao Z, Hu Y, Liang Z, *et al*. Authentication is fundamental for standardization of Chinese medicines. *Planta Med* 2006;72:865–74.
13 Li JW, Vederas JC. Drug discovery and natural products: end of an era or an endless frontier? *Science* [Research Support, Non-U.S. Gov't Review] 2009;325:161–5.
14 Chen ST, Dou J, Temple R, *et al*. New therapies from old medicines. *Nat Biotechnol* 2008;26:1077–83.
15 McCray AT, Ide NC. Design and implementation of a national clinical trials registry. *J Am Med Inform Assoc* 2000;7:313–23.
16 Hegnauer R. Chemical characters in plant taxonomy: some possibilities and limitations. *Pure Appl Chem* 1967;14:173–87.
17 Saslis-Lagoudakis CH, Klitgaard BB, Forest F, *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:e22275.
18 Fairbrothers DE MT, Scogin RL, Turner BL. Bases of angiosperm phylogeny—chemotaxonomy. *Ann Mo Bot Gard* 1975;62:765–800.

19 Harborne JTB. *Plant chemosystematics*. London: Academic Press, 1984:562.
20 Ososki AL, Kennelly EJ. Phytoestrogens: a review of the present state of research. *Phytother Res* [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review] 2003;17:845–69.
21 Giam X, Bradshaw CJA, Tan HTW, *et al*. Future habitat loss and the conservation of plant biodiversity. *Biol Conserv* 2010;143:1594–602.
22 Hamilton AC. Medicinal plants, conservation and livelihoods. *Biodivers Conserv* 2004:1477–117.
23 Zhu F, Qin C, Tao L, *et al*. Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc Natl Acad Sci USA* [Research Support, Non-U.S. Gov't] 2011;108:12943–8.
24 Lui S, Wei M, Moore R, *et al*. RxNorm: prescription for electronic drug information exchange. *IT Professional* 2005;7:17–23.
25 Roberson E. Medicinal plants at risk—nature's pharmacy, our treasure chest: why we must conserve our natural heritage. 2008. http://wwwbiologicaldiversityorg/publications/papers/Medicinal_Plants_042008_lorespdf (accessed 23 Dec 2012).
26 Hartwell JL. *Plants used against cancer*. Lawrence, Mass: Quarterman Publications, Inc, 1982.
27 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* [Research Support, Non-U.S. Gov't] 2000;28:27–30.
28 Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* [Research Support, Non-U.S. Gov't] 2006;34:D668–72.
29 Afendi FM, Okada T, Yamazaki M, *et al*. KNApSAcK Family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol* 2012;53:e1.
30 Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *J Am Med Inform Assoc* 1997;4:484–500.
31 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32:D267–70.
32 Sayers EW, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 200937:D5–15.
33 Ruggiero M, Gordon D, Bailly N, *et al*. The catalogue of life taxonomic classification, Edition 2, Part A. In: Bisby FA, Roskov YR, Culham A., Orrell TM, Nicolson D, Paglinawan LE, Bailly N, Appeltans W, Kirk PM, Bourgoin T, Baillargeon G, Ouvrard D. Species 2000 & ITIS Catalogue of Life, 3rd February 2012. DVD; Species 2000, Reading, UK, 2009.
34 Bard J. Ontologies: formalising biological knowledge for bioinformatics. *Bioessays* 2003;25:501–6.
35 Olsen G. "Newick's 8:45" Tree format standard. 1990. http://evolutiongeneticswashingtonedu/phylip/newick_dochtml (accessed 23 Jan 2012).
36 IUCN2012. The IUCN Red List of Threatened Species. Version 2012.2. http://wwwiucnredlistorg (accessed 18 Oct 2012).
37 Guralnick R, Hill A. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.] 2009;25:421–8.
38 Stephens CR, Heau JG, Gonzalez C, *et al*. Using biotic interaction networks for prediction in biodiversity and emerging diseases. *PLoS ONE* 2009;4:e5725.
39 Vanherweghem JL, Depierreux M, Tielemans C, *et al*. Rapidly progressive interstitial renal fibrosis in young women: association with slimming regimen including Chinese herbs. *Lancet* [Case Reports] 1993;341:387–91.
40 Johanns ES, van der Kolk LE, van Gemert HM, *et al*. [An epidemic of epileptic seizures after consumption of herbal tea]. *Ned Tijdschr Geneeskd* [Case Reports] 2002;146:813–6.
41 Fenneman J. Understanding synonyms. In: Brian K. *E-Flora BC: Electronic Atlas of the Flora of British Columbia [eflorabcca] Lab for Advanced Spatial Analysis, Department of Geography*. Vancouver; University of British Columbia, 2012.
42 McNaught A. The IUPAC International Chemical Identifier:InChI. *Chem Int (IUPAC)* 2006;28. http://www.iupac.org/publications/ci/2006/2806/2806-pp12-15.pdf (accessed 20 Aug 2012).
43 Jessop DM, Adams SE, Willighagen EL, *et al*. OSCAR4: a flexible architecture for chemical text-mining. *J Cheminform* 2011;3:41.
44 Gao T, Yao H, Song J, *et al*. Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. *J Ethnopharmacol* [Research Support, Non-U.S. Gov't] 2010;130:116–21.
45 CHEMnetBASE. http://www.chemnetbase.com [database on the Internet]. http://www.chemnetbase.com (accessed 27 Jan 2012).
46 CBOL Plant Working Group. A DNA barcode for land plants. *Proc Natl Acad Sci USA* [Research Support, N.I.H., Intramural Research Support, Non-U.S. Gov't] 2009;106:12794–7.
47 Lou SK, Wong KL, Li M, *et al*. An integrated web medicinal materials DNA database: MMDBD (Medicinal Materials DNA Barcode Database). *BMC Genomics* [Research Support, Non-U.S. Gov't] 2010;11:402.
48 Cowan MM. Plant products as antimicrobial agents. *Clin Microbiol Rev* [Review] 1999;12:564–82.
49 Hartwell JL. *Plans used against cancer*. Lawrence, Mass: Quarterman Publications, Inc, 1982.
50 Cragg GM, Newman DJ. Plants as a source of anti-cancer agents. *J Ethnopharmacol* [Review] 2005;100:72–9.
51 Bailey RL, Gahche JJ, Lentino CV, *et al*. Dietary supplement use in the United States, 2003–2006. *J Nutr* 2011;141:261–6.
52 Saba N, Ahmad VU, Ali Z, *et al*. Separation and identification of a new saponin from the flowers of Guaiacum officinale L. *Nat Prod Res* 2010;24:1877–82.
53 Ahmad VU, Perveen S, Bano S. Guaiacin A and B from the Leaves of Guaiacum officinale. *Planta Med* 1989;55:307–8.
54 Fischedick JT, Hazekamp A, Erkelens T, *et al*. Metabolic fingerprinting of Cannabis sativa L., cannabinoids and terpenoids for chemotaxonomic and drug standardization purposes. *Phytochemistry* [Research Support, Non-U.S. Gov't] 2010;71:2058–73.