pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts

JYOTI RANI, AB RAUF SHAH and SRINIVASAN RAMACHANDRAN*

GN Ramachandran Knowledge Centre for Genome Informatics,
CSIR–Institute of Genomics and Integrative Biology,
New Delhi 110 025, India

**Corresponding author (Fax, 91-11-2766-7471; Email, ramu@igib.in)*

The PubMed literature database is a valuable source of information for scientific research. It is rich in biomedical literature with more than 24 million citations. Data-mining of voluminous literature is a challenging task. Although several text-mining algorithms have been developed in recent years with focus on data visualization, they have limitations such as speed, are rigid and are not available in the open source. We have developed an R package, pubmed.mineR, wherein we have combined the advantages of existing algorithms, overcome their limitations, and offer user flexibility and link with other packages in Bioconductor and the Comprehensive R Network (CRAN) in order to expand the user capabilities for executing multifaceted approaches. Three case studies are presented, namely, 'Evolving role of diabetes educators', 'Cancer risk assessment' and 'Dynamic concepts on disease and comorbidity' to illustrate the use of pubmed.mineR. The package generally runs fast with small elapsed times in regular workstations even on large corpus sizes and with compute intensive functions. The pubmed.mineR is available at *http://cran.r-project.org/web/packages/pubmed.mineR*.

[Rani J, Shah AR and Ramachandran S 2015 pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. *J. Biosci.* **40** 671–682] **DOI** 10.1007/s12038-015-9552-2

1. Introduction

Text-mining algorithms are gaining popularity in literature searches and analyses. They offer ease of use and speed for extraction and for analysis of information from large text bodies. PubMed, an online literature database, is a valuable source of information for scientific research spanning the field of biology and medicine. The NCBI PubMed (Canese and Weis 2013) database is very important in biomedical literature (Frisch *et al.* 2009). There are more than 24 million records in PubMed, representing the largest collection of biomedical literature. As the entries in the PubMed grow, extraction of useful information in reasonable time poses significant challenge. Although PubMed offers straightforward and fast search (Frisch *et al.* 2009), the output requires extensive manual reading, which is time consuming. Development of fast algorithms can complement this exercise by offering speed, thereby allowing rapid distillation of the trends and concepts. Such enablement can facilitate rapid reviews, project planning and model or hypothesis building.

To this end, several text-mining algorithms have been developed during the recent years. The main goal of these algorithms has been to aid readers in examining the literature with ease, most notably through visual inspection. Many of these are available as online text-mining algorithms: LitInspector (Frisch *et al.* 2009) offers the extraction of genes and signal transduction pathways; GoPubMed (Delfs *et al.* 2004) offers extraction of GeneOntology terms from

Keywords. Package; PubMed; R, text-mining

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/oct2015/supp/Rani.pdf*

Published online: 29 September 2015

the PubMed abstracts, thereby facilitating relevant subontology; and PolySearch (Cheng *et al.* 2008) offers extraction of associative relationship between human diseases, genes and proteins, single nucleotide polymorphisms (SNPs), subcellular localizations, drugs and metabolites, thereby allowing 63 different types of combinations for search. These algorithms support direct extraction of abstracts or sentences from the PubMed database and of highlight relevant terms such as gene name, disease name, species names, which enable easy reading of the text through visualization. These Web-enabled algorithms, although very valuable, pose limitations such as speed, are somewhat rigid in towing predesigned concept lines and are not available in open source.

Algorithms perform specific tasks step-wise. However, text-mining is multifaceted and depends on users' perspective and experience. Realizing this unique qualification of text-mining, we developed this package with multiple layers of compute functions as part of this embodiment. The principle motivation to develop pubmed.mineR arose from our perceived needs to combine the advantages of existing algorithms, overcome their limitations, offer user flexibility, link with other packages in Bioconductor (Gentleman *et al.* 2004) and The Comprehensive R Network (CRAN), in order to expand the user capabilities for executing multifaceted approaches. Towards this end, a favourable choice of computing environment is R, which is referred to as most versatile statistical computing environment (Feinerer *et al.* 2008). Statistical computation allows pre-processing, clustering, summarization and finding associations (Davi *et al.* 2005).

There are various text-mining methods used in diverse fields such as in linguistic stylometry (Giron *et al.* 2005) and in rankings of documents from search engines (Radlinski and Joachims 2007). The semantic Web allowed the standardized formats to perform semantic operations, which offer flexibility in document exchange (Ingo Feinerer *et al*. 2008).

In this article, we describe an R package, pubmed.mineR, developed with an aim of data-mining of PubMed abstracts using text-mining algorithms for biomedical research purposes. The pubmed.mineR also uses several existing functions from other R packages in order to enable text-mining. Some key facilities are terms extraction and their contexts, gene recognition, association between terms and between genes including cross-associations, and hunting for key evidences of proof of associations or evidences. Three case studies are described to illustrate some potential uses of pubmed.mineR. The work flow of pubmed.mineR is shown in figure 1.

2. Methods

Text-mining consists of the tasks (Cohen and Hunter 2013) described in the following sections.

2.1 *Information retrieval*

The key source here is the PubMed database. Abstracts of articles are saved locally using the PubMed search engine. The two functions in pubmed.mineR – readabs() (for text format) and xmlreadabs() (for XML format) – serve to import either the text file or the XML file into the S4 object of class 'Abstracts'. This S4 data object is the starting point for all subsequent processing.

2.2 *Document classification*

The task of classifying documents into categories is widely used and serves as the first step towards systematic data collection, curation, annotation and to extract patterns satisfying users' queries. This is the most applied task. Meeting the requirements of this task are a wide range of functions: searchabsL(), searchabsT(), combineabs(), removeabs(), cleanabs(), sendabs(), yearwise() and genewise(). The searchabsL() and searchabsT() are most general functions allowing Boolean combinatorial query formulation for classification.

2.3 *Summarization*

Document summarization is the process of constructing automatic summaries of a documentor set of documents to enable users to obtain the gist in short time. The pubmed.mineR package provides the functions printabs(), Find_conclusion(), find_intro_conc_html() and pubtator_function() useful for document summarization. The function Find_conclusion() extracts the conclusion from the abstracts usually from the tail end; the function find_intro_conc_html() fetches the introduction and conclusion from long abstracts or reports the entire short abstract, which together serve as a coupled background, and important conclusions of the abstracts, sub-classified termwise and as a output html file, which can be viewed using any browser. Any number of terms can be used for sub-classification as per users' choice. These can be input in the 'themes' argument of the function.

2.4 *Named entity recognition and normalization*

The most popular named entity is the gene, which is written as gene symbol and gene names. We have currently included the HGNC (Human Gene Nomenclature Committee; Gray *et al.* 2015) official gene symbol list for automated recognition. We have also included automated mapping to Uniprot ID (The Uniprot Consortium 2014). In addition, context searches reporting sentences containing the gene name or symbol enable quick identification. In order to aid

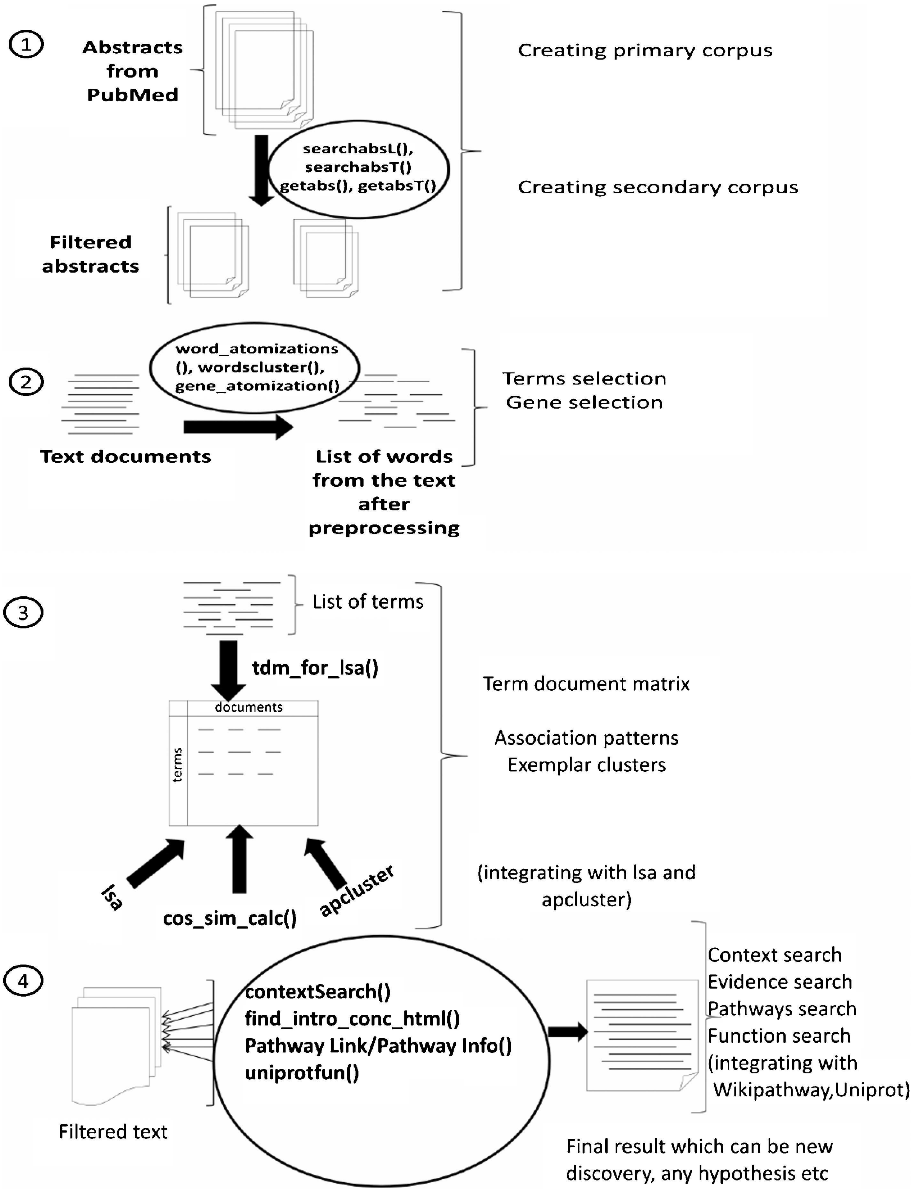

Figure 1. Text-mining facilities of pubmed.mineR. These include classification, summarization and computation. See glossary for explanation of the terminology used.

normalization, mapping to Entrez IDs (Maglott *et al.* 2011) are also provided.

2.5 *Relationships*

A most sought-after relationship is association between genes or cross-association between genes and terms, which could imply their involvement in linked pathways or as common risk factors in the case of diseases. Two-way associations between genes and terms have the potential to reveal the underlying connection between genotype and phenotype. These associations are computed by generating a term document matrix (tdm) through the function tdm_for_lsa(), which reports the raw occurrence frequencies of a given term in each abstract. Subsequently, the association values can be computed either through semantic analysis using lsa (Wild 2007) package in R, or using the function cos_sim_calc() for computing the cosine similarity and, as required, boot strap analysis can also be carried out using the function cos_sim_calc_boot().

2.6 *Linguistic structure*

The linguistic structure in this package has been tackled at two levels: (1) sentence tokenization and (2) word tokenization. The function word_atomizations() tokenizes text into words and reports them in descending order of their occurrence frequencies. Common English words are automatically removed. The top ranking words are high-occurrence-frequency terms. A term is considered to be important on the basis of its occurrence frequency (Feinerer *et al.* 2008). The principle used here is that the smallest entity of the language is a 'word', and several words are connected to construct a sentence. Both words and sentences have meanings at different levels. A very popular procedure used in many text-mining approaches is stemming, which serves to reduce the words to a minimal word length.

Stemming is very attractive, but we observed that applying such reductionist approach may lead to loss of intended usages and to difficulties of mining newly constructed words based on the original word. In addition, scientific texts use other alphabets (for example, Greek alphabets) and hyphens and slashes that may be required by users as such. Therefore, we offer a set of functions that serve to cluster words based on their similarity in alphabet sequence with provision for varying the stringencies. The clusters can then be viewed directly and the appropriate terms selected. The sentence tokenization is used as an intermediate step in several functions, such as in summarization function. The functions used to process linguistic structure are word_atomizations(), SentenceToken(), cluster_words(), wordscluster(), wordsclusterview(), whichcluster() and get_original_term().

2.7 *Data visualization using Cytoscape*

We used Cytoscape version 2.8 (Saito *et al.* 2012) for preparing the association networks from the cosine similarity data between the terms. We used the option degree sorted circle layout with thickness of edges scaled to the association values. We used gray scale for edges and colours for nodes. Green colour was mapped to nodes with low degree, and red colour was mapped to nodes with high degree, in the green-to-red colour band.

3. Results and discussion

3.1 *Case study 1: Evolution in the role of diabetes educators*

In this case study we considered the review article by Scott Drab entitled 'The evolving role of the diabetes educator' (Scott Drab 2013), which describes the impact of diabetes educators (DEs) and diabetes educational services on patient health. The author had selected relevant PubMed abstracts by using the following keywords: 'diabetes educator', 'diabetes education', 'self-management education', 'efficacy/effectiveness', 'outcomes', 'patient-centered medical home', 'barriers', 'referrals' and 'reimbursement'. We used our package pubmed.mineR to extract similar information with a different start point – abstracts until 15 July 2013 from PubMed with the keyword 'diabetes', and the work flow is shown in figure 2a.

The size of each of secondary corpus is shown in table 1. It is evident that the numbers of abstracts are in proportion to the type of the term in regards to its usage – general or specific. The size of the secondary corpus under the term 'outcomes' tops the list, whereas the corresponding size under the term 'patient-centered medical home' is the least. The data matrix for the number of abstracts for each term with co-occurrences of the remaining 9 terms in each set was generated. This matrix is displayed in figure 3. The resulting abstracts constitute tertiary corpora. The matrix cells with same terms (diagonal elements) were set to NULL to exclude data on self-searches. A pertinent question can be asked here – which terms cluster together and how many clusters are formed based on their usage in the literature?

Affinity propagation clustering, a recent technique, is becoming popular (Frey and Dueck 2007). In this approach, in addition to forming clusters, exemplars of clusters are identified among the members. Exemplar member has linguist similarity to all other members in the cluster. The data matrix (figure 3) was input to apcluster (Bodenhofer *et al.* 2011), an R package for affinity propagation clustering, which is a low-error and high-speed algorithm. The results are shown in figure 4. Two representative clusters were formed with exemplars 'efficacy' and 'self-management education'. It is evident that the terms 'outcomes', 'efficacy' and 'effectiveness' are clustered under the exemplar 'efficacy', which are assessment terms. The terms 'diabetes education', 'diabetes educator', 'self-management education', cluster together under the exemplar 'self-management education' as belonging to the same theme of activity terms. It was noteworthy to find the terms 'referral', 'reimbursement', 'patient-centered medical home' and 'barriers' also group in this cluster.

The first 20 top ranking terms according to their frequencies of occurrence in each secondary corpus are shown in table 2. The terms common to all secondary corpora were removed because these are general terms. The results show that the top ranking terms belong to the subject category of the classified abstracts and, in addition, present terms of high importance in the relevant field. For example, in the secondary corpus 'diabetes educators', the terms with high frequency were 'educators', 'education', 'type', 'insulin', 'study', 'results', 'management', 'control' and 'glucose', thereby indicating that the diabetes educators' goals are to control the disease inpatients health

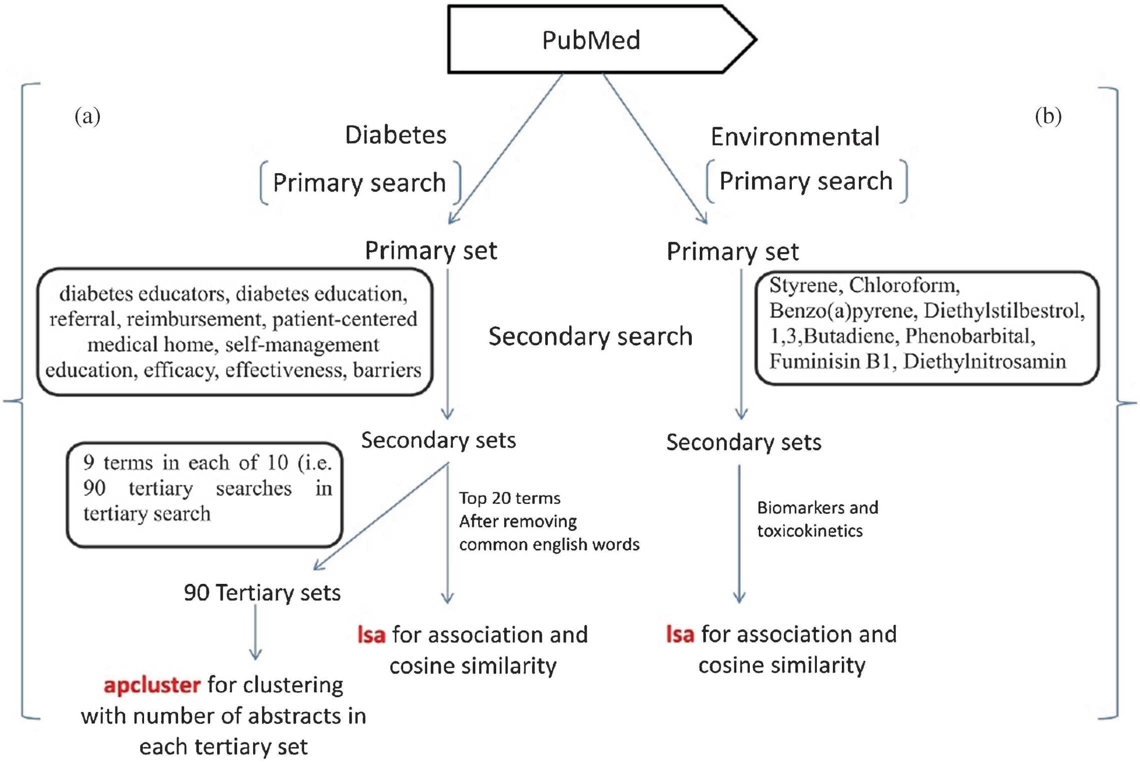
PubMed
(a)
(b)
Diabetes
Primary search
Environmental
Primary search
Primary set
Primary set
diabetes educators, diabetes education, referral, reimbursement, patient-centered medical home, self-management education, efficacy, effectiveness, barriers
Secondary search
Styrene, Chloroform, Benzo(a)pyrene, Diethylstilbestrol, 1,3,Butadiene, Phenobarbital, Fuminisin B1, Diethylnitrosamin
Secondary sets
Secondary sets
9 terms in each of 10 (i.e. 90 tertiary searches in tertiary search
Top 20 terms
After removing
common english words
Biomarkers and
toxicokinetics
90 Tertiary sets
lsa for association and cosine similarity
lsa for association and cosine similarity
apcluster for clustering with number of abstracts in each tertiary set

Figure 2. Flowcharts displaying the analysis steps in the case studies (**a**) ‘Evolution in the role of diabetes educators’ and (**b**) ‘Cancer risk assessment’.

through management. The term ranked next to the top most term in each corpus offers various clues about the profile, such as education, which is the primary goal of educators.

Similarly, ‘study’ follows ‘control’ in the ‘diabetes education’ category, ‘insulin’ follows ‘barriers’ in ‘barriers’ category, ‘risk’ follows close to ‘outcomes’ in the ‘outcomes’ category, ‘control’ follows ‘self-management’ in the ‘self-management’ category, ‘efficacy’ follows ‘treatment’ in the ‘efficacy’ category, ‘effectiveness’ follows ‘treatment’ in the ‘effectiveness’ category, ‘disease’ follows ‘referral’ in the ‘referral’ category and ‘costs’ follows ‘reimbursement’ in the ‘reimbursement’ category. It is apparent that treatment is intimately connected to ‘efficacy’ and ‘effectiveness’, and ‘costs’ are an immediate concern for ‘reimbursement’. A major goal of self-management is to control the disease. The relevancy of ‘insulin’ as THE immediate term to ‘barriers’ refers to ‘insulin’ treatment and ‘risk’ as immediate terms to ‘outcomes’ illustrates the intimate link between health outcome and risks of an undertaken activity. The remaining terms within each secondary corpus elaborate the context of these intimate connections.

It is noteworthy that in the secondary corpus ‘referral’, the ‘age’ of the patient and ‘risk’ constitute the relevant context, which are absent among top terms in other secondary corpora.

Next, we examined the association strength between these terms (table 2) using the cosine similarity function of ‘lsa’ package, an R package for latent semantic analysis. The association values output from the associate function were

Table 1. Number of abstracts for the terms used by Drab (2013) from the primary diabetes corpus of 3,83,270 abstracts

Terms	Number of abstracts
Outcomes	19484
Efficacy	12742
Effectiveness	6702
Self-management education	2205
Referral	2183
Barriers	1905
Diabetes education	1386
Diabetes educator	600
Reimbursement	461
Patient-centred medical home	40

	diabetes educator	diabetes education	efficacy	effectiveness	referral	reimbursement	patient-centred medical home	self-management education	outcomes	barriers
diabetes educator	Null	142	27	36	29	12	1	44	111	51
diabetes education	142	Null	107	130	31	23	1	79	265	87
efficacy	27	107	Null	851	80	31	2	43	1532	158
effectiveness	36	130	851	Null	115	63	2	55	1479	124
referral	29	31	80	115	Null	16	3	8	350	54
reimbursement	12	23	31	43	16	Null	3	19	74	27
patient-centred medical home	1	1	2	2	3	3	Null	0	10	1
self-management education	44	79	43	56	8	19	0	Null	107	34
outcomes	111	265	1532	1479	358	74	10	107	Null	360
barriers	51	87	158	129	54	27	1	34	360	Null

Figure 3. Number of abstracts for a given pairwise combination of terms in the tertiary sets. Each cell in the matrix reports the number of abstracts for the given pairwise combination of terms corresponding to the cell. Note that the diagonal representing self-match has been set to 'NULL' in this matrix. In the case of cross-combination between 'self-management education' and 'patient-centered medical home', there are no abstracts with co-occurrence of these two terms.

input to Cytoscape (Saito *et al.* 2012) for graphic visualization of results. The result for each individual secondary corpus is shown in supplementary figure 1.

Text classified under 'barriers' (supplementary figure 1a) had 4 term nodes, all with equally high degree of 6 and with edges of high association (0.9). These nodal terms were 'management', 'risk', 'study' and 'barriers'. We refer to these terms collectively as 'terrace' terms. As observed with the occurrence frequencies, the 'insulin' node is connected to the node 'treatment'. It is evident that the secondary corpus on barriers focuses on the study of barriers in relation to risk and management. In order to extract further details on these 'terrace' terms, we input the terms to the contextSearch() function and observed examples of some financial and geographical barriers such as poorly written literacy, limited access to diabetes education, low effectiveness, complexity to health care and poor communication between consumers and healthcare professionals. Future educational interventions need to be focused on the risk factors for kidney disease.

The network of association between top ranking terms under 'diabetes education' (supplementary figure 1b) had 1 apex concept node 'results' with a degree of 8 with high association edges. Context search revealed that various studies were conducted to assess the relationship between educators and diabetes patients, and diabetes educators need to

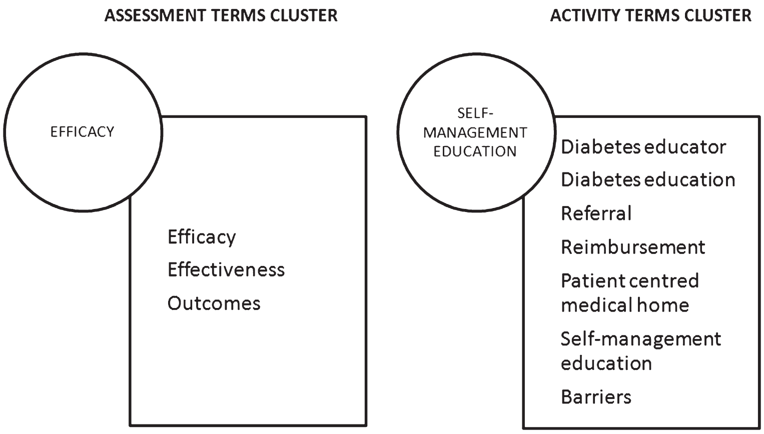

Figure 4. Results of 'apcluster' on the basis of Euclidean distance showing two exemplar clusters 'Efficacy' and 'Self-management education'. Terms clustering together on the basis of their usage in the literature are shown in their respective clusters.

Table 2. Top terms, arranged according to their frequency of occurrence, within each secondary corpus obtained through 'word' tokenization†

Diabetes educators	Diabetes education	Barriers	Outcomes	Patient-centered medical home	Self-management	Efficacy	Effectiveness	Referral	Reimbursement
educators (628)	control (1003)	barriers (2731)	outcomes (27159)	medical (77)	management (681)	treatment (14123)	treatment (6356)	referral (2510)	reimbursement (566)
education (567)	study (922)	insulin (1471)	risk (20475)	primary (47)	self-management (554)	efficacy (14036)	effectiveness (6345)	disease (1978)	costs (330)
type (368)	type (864)	type (1276)	study (17007)	patient-centered (46)	control (255)	educators (13447)	insulin (5835)	study (1939)	results (325)
insulin (355)	program (821)	study (1266)	results (15664)	practice, home (44)	study (249)	study (10603)	study (5558)	results (1725)	disease (319)
study (330)	results (800)	results (1221)	program (15615)	quality (43)	type (241)	type (9763)	results (5240)	diabetic (1738)	study (316)
results (316)	knowledge (683)	control (1177)	associated (12795)	pcmh (36)	intervention (225)	diabetic (9009)	glucose (4998)	risk (1630)	risk (310)
management (266)	intervention (586)	disease (1017)	clinical (12309)	measures (33)	results (221)	results (8694)	type (4723)	age (1306)	data (282)
program (264)		treatment (999)	treatment (11892)	services, model (32)	program (218)	glucose (8366)	risk (4430)	treatment (1253)	medical (261)
control (260)		risk (914)		disease (31)	participants (196)	therapy (7883)	clinical (4133)	years (1204)	treatment (256)
glucose (243)		methods (897)		clinical, chronic (30)	outcomes (194)	clinical (7556)	control (4091)		type (253)
		management (845)			dsme (166)	control (7390)			
						risk (6521)			

† Word tokenization was carried out using the function word_atomizations(). Column titles represent the titles of secondary corpora. Top 20 terms were selected from each secondary corpus and terms common to all the secondary corpora were removed because these are general terms. The numerical values in the parentheses beside each term represent total frequencies of occurrence of the term in the corresponding secondary corpus. If two terms had same frequency of occurrence they were clubbed in the same cell.

more aggressively shape their role and promote the provision of their services for adequate education to diabetic patients. The knowledge term appears in the 'diabetes education' network, whereas the term 'glucose' appears in the 'diabetes educator' network (supplementary figure 1c).

In the case of efficacy network (supplementary figure 1d), 5 terrace concept nodes each with 6 degree of high association are 'study', 'results', 'control', 'clinical' and 'efficacy'. The contextSearch() with these terrace terms revealed that the efficacy of a treatment or medical procedure is evaluated in the clinical studies towards control of the disease.

The network of association in 'reimbursement' corpus (supplementary figure 1e) had 2 concept nodes each with 6 degree of high association. These are 'results' and 'data'. Using these two terms as input for contextSearch() from the 'reimbursement' abstracts showed that various surveys are discussed with their results in the output context of given terms showing evolution for feasibility of telemedicine management of diabetes patients which are also clinically effective. The 'costs' term is strongly associated with 'data', 'medical', 'treatment' and 'disease'.

In the case of effectiveness network (supplementary figure 1f), 1 apex concept node clinical with 4 degree of high association appears. The contextSearch() revealed that effectiveness of any program is evaluated in terms of clinical outcomes.

The patient-centered medical home association network (supplementary figure 1g) had 1 concept term 'patient-centered' with 6 degree of association. The output context of this term defines the patient-centered medical home, which is a team-based care with involvement of physicians and nurses to improve quality of healthcare. Diabetes standards of health care are improved in patient-centered medical home by visit of pharmacists. Results also suggest that integration of registered nurse and certified diabetes educators in a patient-centered medical home improves clinical outcomes and is cost-effective.

The self-management network (supplementary figure 1h) had 1 apex concept term node 'study' with 7 degree of high association. The contextSearch() output reveals that various studies are conducted to evaluate user satisfaction with diabetes self-management education (DSME) programs. It was observed that following DSME, participants maintained improved health parameters such as glycaemia, weight and lipid profile.

The outcomes network (supplementary figure 1i) had 3 concept nodes each with 5 degrees of high association. These are 'outcomes', 'clinical' and 'treatment' and the context search with these terms revealed that a key focus is to evolve strategies to optimize type 2 diabetes care management. Treatment outcomes have been evolving over the past several years, which is indicator of the evolution in the role of diabetes educators.

The association network on referral (supplementary figure 1j) had 1 concept term node 'year' with 8 degree of high association, where 'year' or 'years' refer to the age of the patient or duration of the treatment.

3.2 *Case study 2: Cancer risk assessment using text-mining*

Cancer risk assessment (CRA) refers to assessment of risks causing cancer by evaluating an individual's chance of developing cancer either based on personal or family history. Korhonen *et al.* (2009) carried out the study of CRA considering the available CRA guidelines of United State (US) Environmental Protection Agency (EPA). Through text-mining and experts' consultation, Korhonen *et al.* have generated a taxonomy tree.

We prepared the primary corpus from the PubMed using 'environmental' as the search keyword with an advanced filter on the dates of publication ranging from 1/1/2009 to 10/2/2014. This work-low is shown in figure 2b. The corpus with 3, 65,598 numbers of abstracts was processed, and eight secondary sets of search partitions were made using the names of eight chemicals (Korhonen *et al.* 2009). The results are shown in table 3. The corpus 'Styrene' was the largest and the corpus 'diethylnitrosamine' was the smallest.

Subsequently, words were extracted from the secondary corpora and screened to enlist the Medical Subject Heading (MeSH) terms (2014) (*http://www.nlm.nih.gov/mesh/*).

Eight sets of term-document matrices were generated. Associations between the terms were measured. Subsequently, 'proof of evidence' was carried out for deep mining.

As an example we describe the results for 'biomarkers' and 'toxicokinetics' adopted from Korhonen's work to explain their associated terms with proof of evidence. The terms associated with 'toxicokinetics' and 'biomarkers' with threshold 0.7 are shown in table 4. The associated terms reveal the subject context against the original term used by

Table 3. Carcinogenic chemicals and their corresponding number of abstracts and the number of MeSH terms†

Chemicals Name	Number of abstracts	Number of MeSH terms in abstracts
Styrene	1074	1927
Chlororform	510	1533
Benzo(a)pyrene	382	1094
Diethylstilbestrol	114	659
1,3,Butadiene	97	514
Phenobarbital	61	478
Fuminisin B1	37	328
Diethylnitrosamine	17	170

† The chemicals are those used by Korhonen *et al.* (2009) in their work.

Table 4. Selected terms from Korhonen *et al.* (2009) taxonomy and their associated terms and proof of evidence from abstracts

Chemical	Term	Co-occurring terms	Evidence in abstracts†
1,3-butadiene	Biomarkers	Urine, measure, goal, person, questionnaire, fossil, solvents, goals, volunteers, diaries, lifestyle	DHMBA, **measured** in the post-shift **urine** samples, correlated with both pre-shift MHMBA and post-shift DHMBA. **[PMID: 21803]** Information about subjects' **lifestyle** and daily activities were recorded by means of **questionnaires** and activities **diaries. [PMID: 19999825]**
	Toxicokinetics	Xenobiotics. probability	To take into account the impact of human variability on the predicted toxicokinetics, we defined **probability** distribution for key parameters related to **xenobiotics** absorption, distribution, metabolism and excretion. **[PMID: 20122977]**
Benzo(a)pyrene	Toxicokinetics	feces	Blood, liver, kidney, lung, adipose tissue, skin, urine and **feces** were collected at t = 2, 4, 8, 16, 2, 33, 48, 72 h post dosing. In kidney, 3-OHBap kinetics showed a distinct pattern with an initial buildup during the first 8 h post-dosing followed by a gradual elimination (t((1.2)) of 15.6 h) **[PMID: 201866996]**
Chloroform	Biomarkers	2-hexanone, biomonitoring	To illustrate in practice the **biomonitoring** of exposure, several examples of toxics and their associated biomarkers are reviewed benzene, toluene, styrene, polycyclic, aromatic hydrocarbons, chloroform, **2-hexanone** and hydrogen cyanide. **[PMID: 2377663]**
Diethylnitrosamine	Biomarkers	Heating, coconut, catalase, health, risks, temperatures, vegetable, consumption, cooking, oils, fistula, bilirubin, foods, humans, injury, ethanol, hydrocarbons, superoxide, food, therapeutic, analysis, antioxidants, plant, rats, enzymes, exhibits, caspase-3, hepatoma	Repeated **heating** of edible **oils** can generate a number of compounds, including polycyclic aromatic **hydrocarbons** (PAHs), some of which have been reported to have carcinogenic potential. Repeated **heating** of **vegetable oils** at high **temperature**s during cooking is a very common practice. **[PMID: 20687968]**
Diethylestilbestrol	Biomarkers	Skin, 3-methylcholanthrene-deoxyribose, epoxides, hydrocarbon, lymphoma, papillomas, thyroid, quinones, mutations, carcinogenesis, cysteine, metabolism, cancer, etiology, woman, cytochrome, cyst, mutation, benzene.	Elucidation of estrogen carcinogenesis required a few fundamental discoveries made by studying the mechanism of **carcinogenesis** of polycyclic aromatic **hydrocarbons** (PAHs) Significantly higher adduct ratios have been observed in women with breast, **thyroid,** or ovarian cancer. When estrogen metabolism becomes unbalanced. More catechol estrogen quinines are generated, resulting in higher level of estrogen DNA adducts, which can be used as biomarkers of unbalanced estrogen metabolism and, thus **cancer** risk. **[PMID : 23994691]**
Phenobarbital	Toxicokinetics	Females, males, carcinogenesis, immnohistochemistry, aids, risk	However, hepatocellular proliferation, quantified by BrdU**immunohistochemistry**, was the most sensitive indicator of Phenobarbital (PB) exposure with **female** mice more sensitive than **males**, contrary to sex-specific differences in sensitivity to **hepatocarcinogenesis. [PMID:2444942]**

Table 4 (continued)

Chemical	Term	Co-occurring terms	Evidence in abstracts[†]
Styrene	Toxicokinetics	Gonad, isoforms, trout, jaw.	Juvenile brown **trout** (Salmotrutta) were exposed to three different amounts of the beta-**isomer** (low, medium, high) via the food followed by a period in which they were exposed to unfortified food. On days 0, 7, 1, 21, 35, 49, 56, 63, 77, 91, 105 and 133, eight fish from each treatment group were euthanized and liver, plasma, **lower jaw** (i.e. thyroid tissue) and **gonad** were collected and the remaining tissue (whole-fish) were retained. **[PMID:21216340]**
	Biomarkers	Varnish, creatinine	A repeated measurement protocol was applied to measure airbone styrene (StyA) and urinary biomarkers in 10 **varnish** and 8 fiberglass reinforced plastic workers. The mean air concentration of styrene in breathing zone of workers (30.4 ppm) and the mean concentration of urinary metabolites (MA + PGA = 443 ± 44 mg/g **creatinine**) exceeded the threshold limit value (TLV) and biological exposure index (BEI). **[PMID:20117324]**

[†] The co-occurring terms are marked in boldface type as appearing in selected text parts of the corresponding abstracts. The PMID of the reference is also given.

Table 5. Terms from Korhonen *et al.* (2009) taxonomy and newly discovered terms in this work along with their proof of evidence

	Terms used in the Kohonen *et al.* (2009) taxonomy	New terms obtained in this work from toxicokinetics and biomarkers abstracts	Evidence in abstracts	PMID
1.	Urinary elimination	Urinary excretion	The study allowed the identification of the kidney as a retention compartment governing 3-OH BaP a typical **urinary excretion.**	20186696
2.		Xenobiotic metabolism	Several **xenobiotic metabolism** and cell protective pathways were activated at lower dose.	24449422
3.		Dihydroxybutyl-mercepturic acid (DHBMA) and monohydroxy-butenylmercepturic acid (MHBMA)	It is suggested that **DHBMA** is more suitable biomarker of exposure to 1,3-butadiene (BD) in humans	24184043
4.	DNA strand breaks	DNA repair, leukocytes	Biomarkers of early biological effects, 8-hydroxy-2′-deoxyguanosine in **leukocytes** (8-OHdG), DNA-strand breaks, and **DNA-repair** capacity, measured as an increase in gamma ray-induced chromo-some aberrations were significantly higher in traffic policemen than con-trols (p<0.001 for 8-OHdG, p<0.01 for tail length, p<0.001 for olive tail moment, p<0.05 for dicentrics and p<0.01 for deletions)	20627202

The new terms (column 2) are marked in boldface type as appearing in selected text parts of the corresponding abstracts (column 3).

Korhonen *et al.* (2009). This example shows that a short summary of the relevant associated concepts can be mined from the literature in order to reveal the growth of the field. In a step further, we were able to identify new terms used in combination with the original term (table 5). This was revealed through manual examination of the relevant abstracts under each original term.

3.3 *Case study 3: Dynamic concepts on disease and comorbidity*

Here we used three search terms, 'vitiligo', 'comorbidity' and 'homocysteine', and their abstracts were individually processed to extract the terms in these diseases during the years from 2012 to 2014. After pre-processing, we considered only top 100 high-frequency terms from each corpus and the selected terms were used for further analysis. An association between the terms is shown in supplementary figure 2.

In the case of 'vitiligo', the top ranking associated terms varied during 3 years (supplementary figure 2a–c). The term 'vitiligo' has strong association with 'repigmentation', 'patients' and 'dermatology'. 'Skin' has strong association with 'epidermal' and 'melanocytes'. These association patterns reveal the differential focus of the studies carried out. It is noteworthy that in the year 2013 (supplementary figure 2b), additional terms, 'alopecia' and 'psoriasis', appeared, revealing the potential connections between vitiligo and other diseases. 'Alopecia' is strongly associated with 'lesions' and 'depigmentation', whereas 'psorasis' had moderate associations with other terms in the network. 'Stress' has strong association with 'pathogenesis'. In the year 2014 a new term 'phototherapy'appeared among top 100 terms, which was absent in the previous years. 'Phototherapy' is strongly associated with 'patients' and 'dermatology'.

In the case of 'comorbidity', 'diabetes' ranked at the top (supplementary figure 2d–f) and 'women' are highly associated with 'comorbidity' as revealed by studies on comorbidities of diabetes, mostly in women. The terms 'mental', 'depression', 'psychiatric' and 'anxiety' are strongly associated among themselves, as expected, but there are also weaker cross-associations between the terms 'mental', 'women', 'diabetes' and 'psychiatric'. These data suggest a possible connection between the disease and the mental states of the affected patient. In the year 2013, 'cancer' appears among other diseases with weaker association with 'therapy'.

Homocysteine is associated with cardiovascular disease. In the year 2012 (supplementary figure 2g) strong associations are evident between the pairs 'folate' and 'metabolism', 'plasma' and 'concentration', and 'risk' and 'cardiovascular'. In the year 2013 (supplementary figure 2h), new terms 'methylation' and 'methionine' appear along with 'plasma', 'concentrations', 'folic' and 'risk'. But in the year 2014 an additional term 'methylenetetrahydrofolate reductase' (mthfr) appears and is also associated with 'polymorphism' although weakly. 'Stress' (oxidative) also appears and is strongly associated with 'cardiovascular', 'homocysteine' and 'hyperhomocysteinemia'. The strong associations between the terms 'risk', 'plasma' and 'methylation' continues to appear. As expected, 'homocysteine' and 'hyperhomocysteinemia' are always strongly associated. Thus, it is evident that the evolving terms and their new connections in the context of a given disease can be mined.

4. Performance of the package

Sample runs with total time (user and system) elapsed are shown in supplementary figure 3. It is apparent that on a corpus size of 381,233 abstracts belonging to the subject 'diabetes', the time elapsed was about 31.8 s to sub-classify 1410 abstracts belonging to tuberculosis and about 75.5 s to sub-classify 35627 abstracts belonging to obesity. Intermediate between the two, 42.7 s elapsed to sub-classify 19484 abstracts belonging to outcomes. The time elapsed for word tokenization appears to increase linearly with the corpus sizes in the sequence range 1000, 5000, 50,000, 300,000, which increase exponentially. About 12.0 s elapsed to generate the term document matrix of dimensions 10×19484 with 10 terms on a corpus size of 19484 abstracts belonging to the sub-classification 'outcomes'.

These performance sample runs were carried out on a Mac workstation with 2.2 GHz Quad-Core Intel Xeon and 8 GB RAM under Mac OS X version 10.6.8.

In most research fields the corpus sizes are much lower compared to diabetes, which is an intensely pursued research topic. Cancer is another area where one can expect very large corpus sizes. Therefore, the package pubmed.mineR takes small computational times for major computation-intensive tasks. As R has developed parallel computation utilities, we expect to take advantage of this enhancement as well, thereby reducing computational times even further.

5. Conclusion

The pubmed.mineR offers functions that can reveal not only the connections between terms but also new terms appearing among the top ranking terms. Association matrices between the terms can be computed and networks displayed using the Cytoscape. The three case studies with different subjects provided examples demonstrating the versatility of the package. There is no limitation on the size of text documents for analysis. The package pubmed.mineR has most of the basic functions of text-mining without dependency on other packages at the first level of information extraction. The advanced level of clustering is dependent on other packages such as apcluster. The development on R platform allows further advancement through the ever-growing repositories of CRAN and Bioconductor. The performance of the package is generally fast with small times elapsed even for large corpus sizes and computation-intensive functions.

Acknowledgements

The authors thank Smriti Sharma and Inna Mittal for using the package and providing valuable feedback on text-mining and for suggesting improvements to the algorithms.

This work was supported by Council of Scientific and Industrial Research grant BSC0122.

Glossary

Association	A term used to denote 'closeness' in relationship between a pair of terms.
Concept	A word referring to how it works. Examples – diabetes education, self-management, depigmentation, autoimmune.
Corpus	A collection of documents. plural-corpora
Document summarization	A short summary of the document including the most important parts such as brief introduction and conclusion.
Pre-processing	The process of preparing for analysis using mathematical approaches or other search and display utilities. Examples - word tokenization, sentence tokenization.
Term	A word with exact meaning. Examples - patient, vitiligo, diabetes educator.
Term-document matrix	A numerical matrix where terms are in rows and documents are in columns and the cells contain frequencies of occurrence of terms in the documents.
Text classification	Classifying the documents under defined terms.
Themes	Subjects usually defined by terms and preferably non-overlapping.

References

Bodenhofer U, Kothmeier A and Hochreiter S 2011 APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27** 2463–2464

Canese K and Weis S 2013 updated PubMed: The Bibliographic Database; in *The NCBI Handbook [Internet]* 2nd edition

Cheng D, Knox C, Young N, Stothard P, Damaraju S and Wishart DS 2008 PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* **36** 399–405

Cohen KB and Hunter LE 2013 Chapter 16: Text mining for translational bioinformatics. *PLoS Comput. Biol.* **9** e1003044

Davi A, Haughton D, Nasr N, Shah G, Skaletsky M and Spack R 2005 A Review of Two Text-Mining Packages: SAS TextMining and WordStat. *Am. Stat.* **59** 89–103

Delfs R, Doms A, Kozlenkov A and Schroeder M 2004 GoPubMed: ontology-based literature search applied to GeneOntology and PubMed; in *Proceedings of German Bioinformatics Conference* pp 169–178

Drab S 2013 The Evolving Role of Diabetes Educators. *Am. J. Med. Sci.* **345** 307–313

Feinerer I, Hornik K and Meyer D 2008 Text mining infrastructure in R. *J. Stat. Softw.* **25** 1–54

Frey BJ and Dueck D 2007 Clustering by passing messages between data points. *Science* **31** 5972–5976

Frisch M, Klocke B, Haltmeier M and Frech K 2009 LitInspector: literature and signal transduction pathway mining in PubMed abstracts. *Nucleic Acids Res.* **37** 135–140

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, *et al.* 2004 Bioconductor: open software development for computationalbiology and bioinformatics. *Genome Biol.* **5** R80

Giron J, Ginebra J and Riba A 2005 Bayesian analysis of a multinomial sequence and homogeneity of literary style. *Am. Stat.* **59** 19–30

Gray KA, Yates B, Seal RL, Wright MW and Bruford EA 2015 Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* doi:10.1093/nar/gku1071

Korhonen A, Silins I, Sun L and Stenius U 2009 The first step in the development of text mining technology for cancer risk assessment: identifying and organizing scientific evidence in risk assessment literature. *BMC Bioinf.* **10** 303

Maglott D, Ostell J, Pruitt KD and Tatusova T 2011 Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39** D52–D57

Radlinski F and Joachims T 2007 Active exploration for learning rankings from click-through data; in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp 570–579

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, *et al.* 2012 A travel guide to Cytoscape plugins. *Nat. Methods* **9** 1069–1076

The UniProt Consortium 2014 Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42** D191–D198

Wild F 2007 lsa: Latent Semantic Analysis; *R package version 0.63-3*, *http://CRAN.R-project.org/package=lsa*