International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 25(2): 86–100 (2016)
Published online 17 July 2015 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1481

Text mining applications in psychiatry: a systematic literature review

ADELINE ABBE,[1,2] CYRIL GROUIN,[3] PIERRE ZWEIGENBAUM[3] & BRUNO FALISSARD[1,2]

1 Inserm, U669, Paris, France
2 University Paris-Sud and University Paris Descartes, UMR-S0669, Paris, France
3 LIMSI-CNRS, UPR 3251, Orsay, France

Key words
text mining, psychiatry, applications

Correspondence
Adeline Abbé, Maison de Solenn, Unité Inserm U669, 97 Boulevard de Port Royal, 75679 Paris cedex 14, France.
Email: adeline.abbe@u-psud.fr

Received 5 March 2014;
revised 21 January 2015;
accepted 9 April 2015

Abstract

The expansion of biomedical literature is creating the need for efficient tools to keep pace with increasing volumes of information. Text mining (TM) approaches are becoming essential to facilitate the automated extraction of useful biomedical information from unstructured text. We reviewed the applications of TM in psychiatry, and explored its advantages and limitations. A systematic review of the literature was carried out using the CINAHL, Medline, EMBASE, PsycINFO and Cochrane databases. In this review, 1103 papers were screened, and 38 were included as applications of TM in psychiatric research. Using TM and content analysis, we identified four major areas of application: (1) Psychopathology (i.e. observational studies focusing on mental illnesses) (2) the Patient perspective (i.e. patients' thoughts and opinions), (3) Medical records (i.e. safety issues, quality of care and description of treatments), and (4) Medical literature (i.e. identification of new scientific information in the literature). The information sources were qualitative studies, Internet postings, medical records and biomedical literature. Our work demonstrates that TM can contribute to complex research tasks in psychiatry. We discuss the benefits, limits, and further applications of this tool in the future.

Introduction

Text mining (TM) is intended to automatically discover, retrieve, and extract information in a corpus of text, often large, combining approaches involving linguistics, statistics, and computer science. The combination of techniques from natural language processing (NLP), artificial intelligence, information retrieval and data mining help to apprehend the complex analytical processing system of written language (Cohen *et al.*, 2008; Rzhetsky *et al.*, 2009). The first use of TM was mainly outside the medical field, for government intelligence and security agencies to detect terrorist alerts and other security threats. These methods were then widely adapted to other fields, in particular to medicine (Meystre *et al.*, 2008). One of the first biomedical projects was initiated by the University of New York in order to analyse texts written by experts, and consisted in synthesizing the signs and symptoms of patients and identifying possible side-effects of drugs (Sager *et al.*, 1987a, 1987b). Following the technological advances and the development of natural language techniques, the number of publications using TM has more

than doubled in 10 years (Zhu *et al.*, 2013). In the 1990s, Garfield *et al.* (1992) showed how artificial intelligence technology could be used to test theories of psychopathology. In this review, the authors also suggested how researchers and clinicians might begin to think about it as a useful tool in psychiatry. The greatest impact identified was on enhancing both descriptive diagnosis and identifying repetitive themes in content analysis.

New applications of TM have been discussed in recent reviews, specifically in genomics. The abundance of literature and datasets in genetics has led researchers to consider the need for TM tools to identify susceptibility genes potentially involved in genetic diseases, and this could be of particular interest in psychiatric research (Cheng *et al.*, 2008; Yu *et al.*, 2008; Evans and Rzhetsky, 2011).

Secondly, TM tools have steadily increased in accuracy and sophistication, to the point where they are now suitable for widespread application. To achieve new knowledge, TM draws upon contributions of many text analysis components, and on knowledge input from many external disciplines such as computer science, artificial intelligence, management science, machine learning, and statistics (Miner *et al.*, 2012). The basic metric is based on word occurrence in language. The main steps of TM can be described as follows (Miner *et al.*, 2012):

- The creation of a corpus (a collection of documents) by defining inclusion criteria and availability of the data. Data collected includes text documents, HTML files, web postings, and clinical notes.
- A pre-processing step introduces structure to the corpus. This fundamental step is the most significant difference between data mining and TM. The primary purpose is to process unstructured and rough (textual) data in order to extract meaningful information. The second purpose is to convert the corpus into a list of organized elements, i.e. a structured representation of the data. The relationships between key words and documents are characterized by indices, which are relational measures, such as how frequently a given word occurs in a document or how frequently two key words appear in a same sentence.
- Extraction of knowledge: patterns are extracted in the context of a specific problem using knowledge extraction methods (i.e. prediction, clustering, association, trend analysis). Models are developed and validated to see if they actually address the problem and meet the objectives.
- Different approaches have been applied to assess the validity of the results retrieved from TM systems. One straightforward way is a form of face validity assessment, corresponding to the subjective similarity found between the results of the analysis and the perusal of the corpus. Another approach is to compare results to a gold standard. For instance when a TM tool is dedicated to screening depressive disorders in medical records, sensitivity, specificity and other predictive values or receiver operating characteristic (ROC) curve can be estimated from expert ratings of the files (Zweigenbaum *et al.*, 2007).

The general idea of this review is to demonstrate the potential interest of TM when applied to psychiatry. Our present research has two specific objectives: (1) to collect and analyse applications from the studies reviewed in order to assess the benefits and limitations of using TM; and (2) to identify new opportunities for use of TM in psychiatry.

Methods

Selection criteria

The systematic review was conducted independently using the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines, and was consistent with the population, intervention, comparison, outcomes, and time horizon framework (PICOS framework) (Moher *et al.*, 2009). This approach was developed to define research questions. Systematic reviews were used for identification of relevant studies, but were not included in their own right. Full texts were obtained and references lists were reviewed for relevant studies.

Data sources

The systematic literature search to identify applications of TM was performed in the following electronic databases in: the Cochrane Library, MEDLINE (using PubMed platform), Embase, PsycINFO, CINAHL.

Literature search strategy

Papers published up to November 2013 were included in this review. TM applications in any country were included in the search. Electronic database searches were limited to English-language publications. The following exclusion criteria were applied to title/abstract and full text to identify the relevant studies: commentaries and letters, consensus reports, and clinical notes for patients without psychiatric, mental or cognitive disorders.

Hypotheses and limitations

The study selection process comprised the following two phases:

- Level 1 screening: Titles and abstracts of studies identified from electronic databases were independently reviewed for eligibility according to inclusion and exclusion criteria by two researchers (A.A. and B.F.).
- Level 2 screening: Full texts of studies selected at Level 1 were obtained and independently reviewed for eligibility, using the same inclusion and exclusion criteria as in Level 1, by the same two researchers (A.A. and B.F.).

Data extraction

Data were extracted from the full text of articles by two reviewers working independently. Data extracted included the following items: Aim, Research questions, Data collection methods (e.g. questionnaires, interviews, and method of data analysis), Sample characteristics (e.g. participants, age, level of education), Context and setting, Approaches to data analysis (e.g. pre-processing and statistical methods), Key themes, Interest/limitations of TM.

Results

The search yielded 1103 citations. Of these, 895 were ineligible after review of the title and abstract (Figure 1). Thirty-eight studies were included in this review. These studies used interviews, written narratives, and Internet postings from patients with mental disorders, and the biomedical literature.

TM techniques and performance

TM techniques offer varied scope for retrieving relevant information that may be obscured by huge amounts of information. Tables 1–4 summarize applications and TM techniques applied in these psychiatric papers.

Data preparation step

The preparation step has four major components: morphological analysis, syntactic analysis, lexical analysis and dimension reduction.

Morphological analysis helps to delineate words via a phase consisting in cutting into elementary units ("tokenization") followed by normalization by "stemming" or "lemmatization". This analysis automatically cuts a stream of text into words, phrases, symbols, or other meaningful elements:

- The first step of morphological analysis is to remove punctuation and convert text to lowercase.

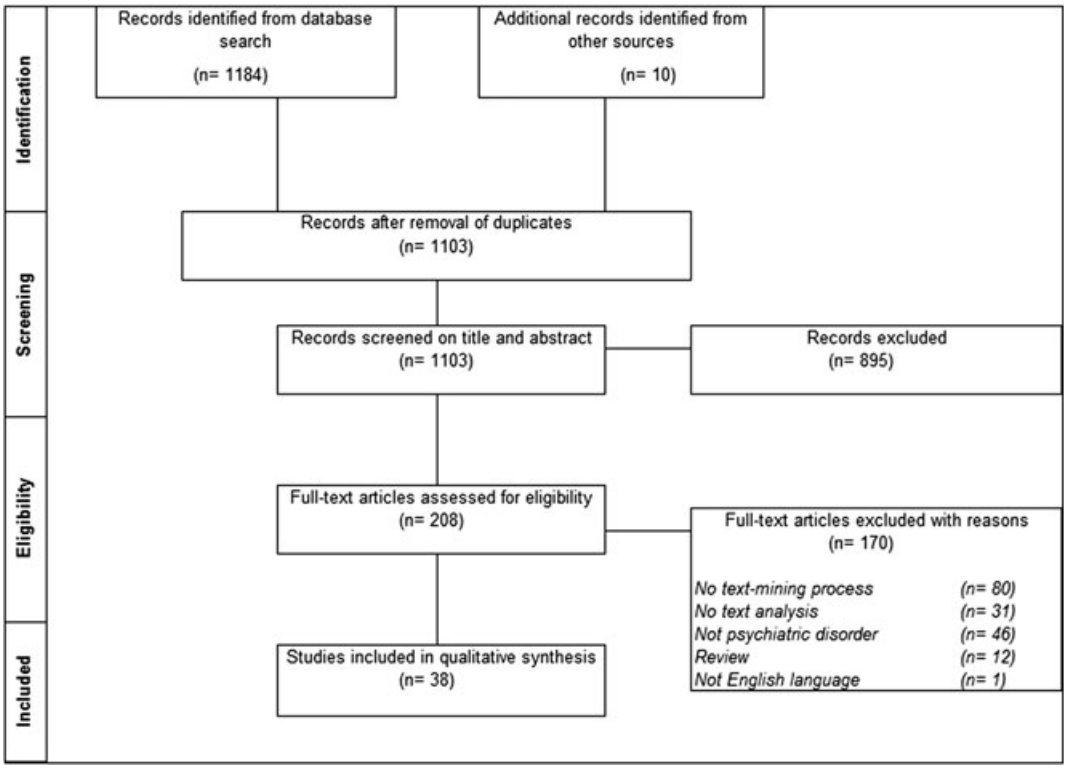

Figure 1. PRISMA flow diagram of the study selection process for application of text mining in psychiatry.

Int. J. Methods Psychiatr. Res. 25(2): 86–100 (2016). DOI: 10.1002/mpr

Table 1. Applications and text-mining methods in psychopathology studies

		Text pre-processing				Data Mining			Text mining software/ program	
Application	Objective	Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association		References
To identify semantic features specific to a disease	To compare written communication of patients with autism spectrum disorders versus control group	Tokenization, Stopword removal	Lemmatization				Cluster analysis, Correspondence analysis		Taltac software	(Bernardi and Tuzzi, 2011)
	To compare social language between normal subjects and patients with autism spectrum disorders				Latent semantic indexing	Classification regression				(Luo *et al.*, 2012)
To identify semantic features specific to a psychological state	To identify emotional content in anxiety	Tokenization		Ontologies		Classification			Tropes	(Piolat and Bannour, 2009)
	To identify anxiety, quality of life and concerns of patients with sleep apnoea	Tokenization[1]				Classification			Tropes, Sphinx	(Veale *et al.*, 2002)
	To examine the impact of incarceration on psychological state of inmates serving long sentences	Tokenization[1]				Classification			ALCESTE	(Yang *et al.*, 2009)
	To investigate the role of different aspects of psychological strain in Chinese rural young suicides	Tokenization[1]				Classification		Correlation tests (Cramer's V)	SPSS Text Analysis for Surveys	(Zhang *et al.*, 2009)

[1]Tokenization was not clearly expressed in the article but suggested in the methods.

Table 2. Applications and text-mining methods to examine patient perspectives

		Text pre-processing				Data mining				
Application	Objective	Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association	Text mining program	References
To evaluate the behaviour of patients	To gain knowledge of the attitudes and behaviours of drug abusers related to the illicit use of pharmaceutical opioids	Tokenization, Stopword removal		Named entity regognition				Co-occurrence	Predose platform	(Cameron *et al.*, 2013)
To identify disorders	To screen for depression in texts	Tokenization[1]			Latent semantic analysis	Logistic regression			Pedesis	(Neuman *et al.*, 2012)
To examine patients' experiences	To screen for post-traumatic stress disorder in patients, using lexical features	Tokenization, Stopword removal			Bag-of-words	Classification		Chi-square test	Text mining approach	(He *et al.*, 2012)
	To understand the process of recovery through sufferers' own words	Tokenization				ANOVA			WordSmith Tools software	(Keski-Rahkonen and Tozzi, 2005)
	To detect causality from online psychiatric texts using inter-sentential language patterns	Tokenization[1]	Parsing					Association rules	Author's text mining module	(Wu *et al.*, 2012)
	To identify information on symptoms experienced, and relationships between symptoms in depression	Tokenization, Stopword removal		Tagging	Vector space model	Classification		Association rules		(Yu and Wu, 2007)
	To identify associations between negative life events and depressive symptoms	Tokenization, Stopword removal		Tagging	Vector space model	Classification		Association rules	DISCOURSE	(Yu *et al.*, 2009)
	To describe the use of language association patterns as features to classify sentences on negative life events	Tokenization, Stopword removal		Tagging	Vector space model	Classification		Association rules	Apriori algorithm	(Yu *et al.*, 2011)

[1]Tokenization was not clearly expressed in the article but suggested in the methods.

Table 3. Applications and text-mining methods in medical records

		Text pre-processing				Data mining				
Application	Objective	Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association	Text mining program	References
To establish safety profiles of a drug	To identify possible adverse events and possible adverse drug events			Tagging, ontology		Classification				(Eriksson *et al.*, 2013)
	To extract physician-asserted drug side-effects	Tokenization	Tagging			Classification			cTAKES	(Sohn *et al.*, 2011)
	To identify adverse drug events	Tokenization	Parsing	Named entity recognition				Chi-square test	MedLEE	(Wang *et al.*, 2009)
To identify genes and pathways involved in a complex disorder	To generate a list of disorders with phenotypes overlapping with SMS	Tokenization[1]						Co-occurrence	MimMiner software	(Girirajan *et al.*, 2009)
	To investigate comorbidity and patient stratification for discovery of overlapping genes	Tokenization			TF-IDF			Correlation	Author's text mining module	(Roque *et al.*, 2011)
To identify relationships among terms in a domain and to build ontologies.	To develop a clinical vocabulary for post-traumatic stress disorder	Tokenization, Stopword removal			Latent semantic indexing	Logistic regression			SAS Enterprise/ Text Miner	(Luther *et al.*, 2011)
To identify relevant findings supporting intermediate diagnosis hypotheses	To extract clinical concepts from psychiatric narrative representing the depressive and manic poles	Tokenization, Stopword removal	Parsing		LSA semantic space	Classification			General Text Parser	(Cohen and Hunter, 2008)
	To classify suicide notes	Tokenization	Parsing	Tagging		Classification, logistic regression, ANOVA			Perl programs, WEKA	(Pestian *et al.*, 2010)

(*Continues*)

Table 3. (Continued)

Application	Objective	Text pre-processing: Morphological analysis	Text pre-processing: Syntax analysis	Text pre-processing: Semantic analysis	Text pre-processing: Dimensionality reduction	Data mining: Supervised learning	Data mining: Unsupervised learning	Data mining: Association	Text mining program	References
	To examine whether patterns identified in diagnostic interviews are associated with diagnostic error in schizophrenia.		Tokenization, Stopword removal			Classification		Kappa	SAS Enterprise software	(Gara *et al.*, 2010)
To improve the accuracy and scope of existing data	To extract Mini Mental State Examination results		Tokenization		Tagging	Classification			Gate	(Cunningham *et al.*, 2013)
	To extract clinical data such as outcomes of antidepressant treatments		Tokenization[1]			Logistic regression, classification, bootstrapping, ANOVA			HiTex platform	(Perlis *et al.*, 2012)
	To determine the number of psychotherapy sessions		Tokenization			Classification			Automated Retrieval Console (ARC)	(Shiner *et al.*, 2012)
	To investigate smoking prevalence and factors influencing this in people receiving mental health care		Tokenization		Tagging	Regression			Gate	(Wu *et al.*, 2013)

[1]Tokenization was not clearly expressed in the article but suggested in the methods.

Int. J. Methods Psychiatr. Res. 25(2): 86–100 (2016). DOI: 10.1002/mpr

Table 4. Applications and text-mining methods for medical literature analysis

		Text pre-processing				Data mining				
Application	Objective	Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association	Text mining program	References
To assess scientific productivity and impact	To identify the top Alzheimer's disease researchers, specific subsets			Disambiguation			Cluster analysis		Thomson-Collexis dashborad	(Sorensen, 2009)
To discover genes implicated in disease	To identify negated relations between genes and disease	Tokenization	Tagging			Classification			Java	(Agarwal *et al.*, 2011)
	To predict autism susceptibility genes	Tokenization[1]		Tagging				Association rules	PolySearch	(Gong *et al.*, 2012)
	To identify genes expressed in Alzheimer's disease	Tokenization[1]						Co-occurrence	LitMiner software	(Liu *et al.*, 2006)
	To identify potentially related genetic disorders			Tagging	Vector space based model			Similarity	Ruby language	(Sarkar, 2012)
To facilitate the annotation of data and literature with terms from ontologies	To uncover patterns and specific trends in TMS for the treatment of depression	Tokenization[1]					Cluster analysis		Matheo-analyzer software	(Dias *et al.*, 2011)
	To retrieve definitions of phobia and germ personality	Tokenization, Stopword removal	Tagging		LSA semantic space	Classification, ANOVA			GALLITO, Matlab	(Jorge-Botana *et al.*, 2009)
	To organize information related to Alzheimer's disease		Chunking		n-gram analysis	Classification			Protegé OWL	(Malhotra *et al.*, 2014)
	To generate text summaries for a set of diseases			Tagging	n-gram analysis			Similarity	SemRep, ROUGE	(Shang *et al.*, 2011)
	To develop an ontology of autism			Tagging		Classification			Protégé OWL	(Tu *et al.*, 2008)
To reduce the burden of updating of reviews	To produce and maintain systematic reviews	Tokenization[1]			Bag-of-words	Classification			LIBSVM 30, Stata	(Wallace *et al.*, 2012)

[1]Tokenization was not clearly expressed in the article but suggested in the method.

- The next step is tokenization, which breaks down a text using the space between words for inflectional languages such as English. This could be difficult for the languages that do not use spaces, or use them inconsistently between words, such as Asian languages.
- Next, a stemming algorithm is applied, reducing a word to its stem or root form without derivational prefixes and suffixes (e.g. both *fishing* and *fished* are reduced to *fish*). It also removes grammatical variants such as present/past and singular/plural.
- Some extremely common words are filtered out because they do not contain important meaning for the search, for instance *the*, *is*, *at*, *which*, and *on* in English, namely *stopwords*.

Morphological analysis extracts terms from the text, but loses the information on relationships among these terms.

Syntax analysis is used to determine the structure linking the different parts of each sentence. Two types of analysis are possible: morphosyntactic labelling, which is an initial step of parsing, and parsing, which determines the relationships among words in a sentence, typically in the form of tree constituents or tree dependency relationships. The identification of parts of speech is established (that is, nouns, verbs, adjectives, and so on) using automatic part-of-speech tagging algorithms. This is done by structuring the language by identifying grammar rules and language conventions, and it contributes to disambiguation.

Syntactic analysis can be complete or partial, or not be used at all. A more advanced form of stemming, known as lemmatization, uses both the context surrounding the word and additional grammatical information such as the part of speech to determine the lemma. For words such as fish, stemming and lemmatization produce the same results. However, for words like meeting, which can serve as either a noun or a verb part, stemming produces the same root meet, but lemmatization produces meet for the verb and maintains meeting if it is the noun. However, syntax is insufficient for understanding meaning fully, and it is often completed by other steps of data preparation. Syntactic analysis can be useful for extracting information (particularly relationships), extraction of terms; it is however not often useful for text categorization.

Semantic analysis provides a real-world clinical interpretation of the sentence, differentiating concepts with a figurative meaning. The semantic rules are developed on the basis of co-occurrence patterns observed in clinical texts. For example, "depressed" and "suicide" would belong to the semantic category "*sign/symptoms*". However, the frequent co-occurrences of both symptoms (<Depressed>, <Suicide>) in the same text is interpreted as a cause–effect relationship. The TM experts divide semantic processing into two types: terminology and ontology (Ananiadou and McNaught, 2006). The main distinction is that:

- if the concepts remains implicit (the human user provides the relationships after the analysis), it is called terminology;
- if the relationships are formalized, the semantic processing is known as ontology.

The following two studies illustrate the two types of semantic analysis. In one study, emotional concepts relative to suicide were allocated to different classes of emotional states based on PubMed queries, so that this can be considered as an ontology (Pestian *et al.*, 2010). In another study, the semantic characteristic relating to affection, emotion, and affective state was searched for in dictionary definitions including synonyms and antonyms. After this step, the authors decided which terms could be included in the following classes: emotional lexicon (such as anger and gaiety) and pleasant or unpleasant psychological states (such as depression and euphoria). Ontologies based on semantic analysis allow text to be mined for interpretable information about biomedical concepts, as opposed to simple correlations discovered by mining textual data using statistical information about co-occurrences of biomedical terms.

The last component of the data preparation is its representation. The aim of this step is to represent a list and the frequency of the terms used in each document.

- Firstly, it creates a structured representation of the data, often referred to as the term–document matrix (TDM). Each row represents a document and each column shows the terms occurring. In this automated process, the previously detected term variants are grouped together into an equivalent terminology (Ananiadou *et al.*, 2010). The relationships between the terms and the documents are characterized by relational measures, such as how frequently a given term occurs in a document.
- Secondly, several methods can be used in order to obtain a more consistent term–document matrix. Raw frequency values are normalized using log frequencies, binary frequencies, or inverse document frequencies. Then singular value decomposition (SVD) can be used in latent semantic analysis (LSA) to find the underlying meaning of terms in the various documents (Han *et al.*, 2011). The method, also called latent semantic indexing (LSI), is widely used as a dimensional-reduction technique to compare similar concepts/topics in a collection of terms.

<NI>The Words*documents matrix thus obtained provides access to all words in each document. Words in a document can be used for information retrieval tasks, by searching texts that are relevant to information expressed in a query. The method is typically based on a measure of similarity between the textual content of the request (words it contains) and the texts in the corpus.

Statistical analysis

In addition to basic descriptive statistics (words counts) and to singular value decomposition of the term*document matrix, many statistical methods can be applied to structured data obtained from the data preparation step. Association analysis is used to automatically identify associations among treatments, genes and diseases. Association rules, correlation tests, co-occurrences and similarity indexes provide association measures. Supervised predictive TM algorithms are used to classify texts. They include the naive Bayes classification, decision trees, logistic models, support vector machines, bootstrap procedures, regression models, and analysis of variance (ANOVA). Unsupervised learning can also be used to identify clusters in the corpus.

Patterns observed in approaches adopted by different publications

We also attempted to automatically identify hidden patterns in clinical studies included in our systematic literature review. Titles and abstracts of each study selected were analysed using a TM approach implemented in the R package. First, we created a dataset with the frequencies of words for each study. Then we applied hierarchical clustering analysis to find similarities between key words. Finally, clustering analysis yielded four distinct types or groups of literature topic, shown in Figure 2.

- Gene expression profiling

In these papers, TM tools helped to extract gene expression profiles for various mental disorders from the literature.

- Representation of psychiatric illnesses

The frequency of the association between certain words and psychiatric illness is measured.

- Exploration of drugs and illness using TM

The TM approach was especially used to extract information concerning drugs and mental illness (side effects, which drug for which disorders, etc.).

- Methods in TM

The improvement of TM tools is a growing topic of discussion. These publications focused on the extraction of relationships based on sentences in biomedical literature and electronic medical data. The articles illustrate innovation in TM using a psychiatric disorder or autism as an example.

Fields of application

In addition to TM of publication abstracts, a content analysis of the papers themselves was performed. The difference with the previous section is the process of classification. Here content analysis is subjective by essence, while previous patterns were obtained from automatic text classification. Both approaches are interesting and complementary, they enable a so-called "triangulation" of the analysis. A consensus

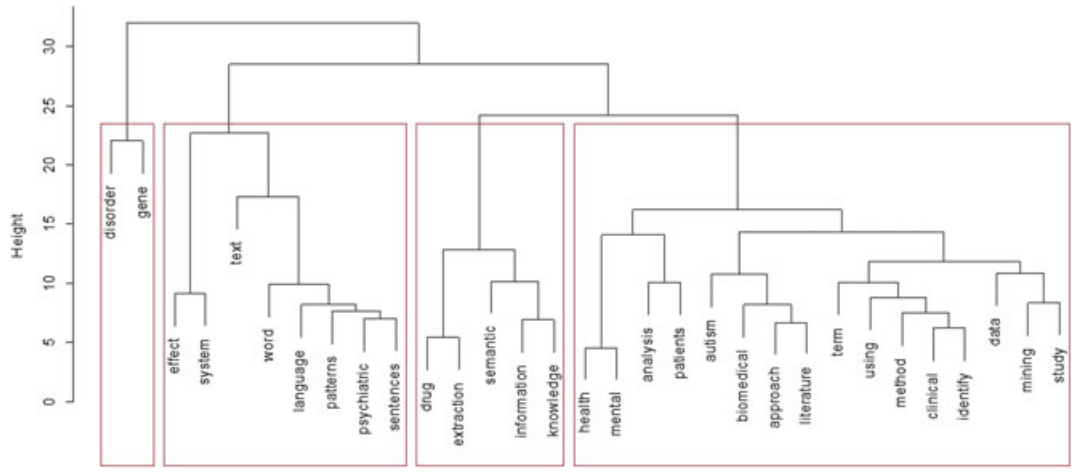

Figure 2. Cluster analysis results with topics of abstracts included.

meeting was organized to homogenize the findings. Four main themes were identified from the 38 studies included: (1) Psychopathology (i.e. the study of mental disorders or mental distress); (2) The patient perspective (patients' thoughts, feelings and behaviours), (3) Medical records (safety issues, quality of care, description of treatments); (4) Medical literature (*ontology* – mapping terms with domain-specific concepts, or biomarkers; *experts* – determining each scientist's main line of investigation; *uncovering hidden topics* – looking for thematic divisions in the domain).

Psychopathology

In these papers, the corpus comprised written observations or patient narratives. Participants' responses were collected from interviews or self-report questionnaires. Six studies used TM to identify semantic characteristics specific to a psychological state or illness, and are shown in Table 1. Two studies compared social language features between normal subjects and patients with autism spectrum disorders, using supervised or unsupervised statistical methods (Bernardi and Tuzzi, 2011; Luo *et al.*, 2012). The other studies examined anxiety and independent factors that influence illness or the psychological state of patients (e.g. related to behaviours, thoughts, emotions, or identity-related factors). Worries and anxieties related to the patient's condition were extracted using classification models. For example, factors predictive of risk for suicide were identified in China on the basis of words used in suicide notes (Zhang *et al.*, 2009). Further to this, the impact of imprisonment on the psychological state of prisoners has been studied in France using classification models (Yang *et al.*, 2009).

The patient perspective

This theme concerns the thoughts, feelings, and behaviours of patients. Internet has become a source of information, support, treatment, and prevention for patients. An increasing number of patients interact online and share their experiences of illness, diseases and therapies. Patients post messages in discussion groups on websites. These messages are considered as the patient's view. As described in Table 2, eight studies explored patients' experiences expressed in their messages concerning the recovery process, negative life events in relation to symptoms, cause–effect relationships, and behaviours of drug abusers. In this category, one study aimed to understand the process of recovery through the sufferers' own words (Keski-Rahkonen and Tozzi, 2005). In addition, three studies focused on negative or stressful life events described by patients on a virtual psychiatric services website. In these studies, Yu *et al.* (2008) used these messages to identify the associations between events and depressive episodes. Two studies conducted further investigation to automatically detect depressive symptoms from questions addressed by patients to Mentalhelp.net and PsychPark.org (Neuman *et al.*, 2012; Wu *et al.*, 2012). Screening natural language in texts is challenging, particularly on the Internet. The language is fragmentary, with typographical errors, often without punctuation, and sometimes incoherent.

Medical records

Patient information is increasingly captured in electronic medical records (EMRs) by caregivers. Records include medical history, treatments, and laboratory and other test results. However, this unstructured textual data is unwieldy to analyse. Thirteen studies investigated whether TM can explore EMRs to detect safety issues, symptoms, comorbidities, patient subgroups and characteristics of therapy (Table 3). TM was used to capture patients' medication histories and response to drugs. Supervised models applied to electronic medical records helped to identify predictive features including drug side effects, treatment-resistant depression, symptoms, and psychotherapy sessions received. Further to this, comorbidities, and drug side-effects were analysed to identify overlapping genes using unsupervised learning models. Roque *et al.* (2011) demonstrated how records from psychiatric hospital enable the identification of correlations between diseases. Medical records were also used to identify relevant findings supporting diagnosis hypotheses for schizophrenia and mood spectrum disorders. Cohen and Hunter (2008) extracted clinical concepts from psychiatric narrative by depressive and manic patients. Information held in structured fields could be usefully supplemented by open-text information such as smoking status, examination results, number of psychotherapy sessions and outcomes of antidepressant treatment.

Medical literature

The biomedical literature is currently expanding at a rate of several thousand articles per week. The exploration of this source is more feasible using TM methods. Eleven publications provide practical examples of data retrieved from the literature (Table 4). Three studies developed a clinical terminology to map terms for specific concepts in depression, phobia and autism using association analysis. In addition, TM of PubMed abstracts was employed to identify susceptibility genes in Smith–Magenis Syndrome, autism and Alzheimer's disease. TM techniques were also used to identify expert researchers in a scientific domain,

to uncover patterns and specific trends within the literature and to update systematic reviews.

Discussion

In this systematic review of the literature on TM applied to psychiatry, we found that the techniques used and the topics under study were heterogeneous. From a technical point of view TM always began by reduction, simplification, coding of a given corpus. Then exploratory multidimensional analyses are usually used to process the data obtained. In the most sophisticated approaches, semantic models can also be estimated and tested. Concerning the topics tackled, they differed widely, ranging from genetics, characterization of the patient perspective, automatic detection of symptom patterns to treatment side effects.

Previously, two reviews of TM applications have made similar findings in cancer (Korhonen *et al.*, 2012; Zhu *et al.*, 2013). Zhu *et al.* (2013) concluded that TM is useful to extract new information from qualitative studies, medical records and biomedical literature. The authors encouraged the application of biomedical TM technologies in the development of personalized medicine. In particular, many risk factors associated with disease remain to be explored, such as gender, age, race and environment. Similarly, Korhonen *et al.* (2012) demonstrated how TM could be used to promote knowledge in cancer risk assessment. TM has obvious advantages. It enables systematic, automatic searches and textual data processing. Content analysis has been used to analyse textual data, but this valuable approach is limited to fairly small corpuses and it is highly dependent on the skills of the professional performing the content analysis, and on his or her ability to allow for his/her own subjectivity. Indeed a TM algorithm is not liable to subjectivity, and while the choice of the algorithm and the interpretation of the results are never totally neutral, this is also true of all statistical analyses. Of course, the increasing volume of publications of all sorts, and more generally of textual data in medicine, makes TM a fast-growing tool. However, TM has also several limitations. First, a large corpus is necessary to obtain robust results. In addition, the format of many texts limits the availability of documents that can be mined, for instance publications stored as images are unsuitable. The system also fails to cope with homonymy and polysemy, and disambiguation of different meanings according to context. The algorithms used for concept-based processing are another potential source of bias. The investigator's own subjective interpretations can influence the quality of summarization labels. To minimize the subjectivity bias, some authors used techniques and algorithms capable of summarizing high-level semantic content in unstructured text. Finally, the lack of transparency in the use of TM systems has been criticized. TM is viewed as a black box receiving an input of documents, and this can discourage researchers. Finally the reliability of results obtained by TM is rarely discussed, and this was mostly the case in the studies included in this review. This can be explained by the fact that TM is exploratory in nature, and also by the fact that the notion of reliability is itself vague, at least when no gold standard can be envisaged.

Concerning the present systematic review, it was performed according to PRISMA guidelines using an electronic search of all studies in English, with no limits on publication date. It was restricted to studies using a system that automatically extracts and converts unstructured text documents into data for analysis. Semi-automatic tools such as NVivo for content analysis are increasingly used in qualitative research (Ranney *et al.*, 2014). However, these applications are not included in this review, since they do not fully automate all the steps of the analysis.

Our review highlights the opportunity to give a voice directly to patients using TM. Until now, only patient-reported outcome instruments offered the possibility of collecting patient perspectives. However, closed questions are the most commonly used in patient-reported outcomes and this may lead the respondents in certain directions. In addition, TM can discover new variables from the clinical experiences reported directly by the patients. Patients talk freely about their experiences of treatment, which can provide extra information for the standard descriptions of drugs.

The use of NLP systems in medicine is not easy because it processes two types of vocabulary (patient versus physician). In psychiatry, an additional challenge appears in so far as psychiatric disorders can have an impact on language, and reinforce the need for the *ad hoc* tasks of NLP. Not only will the patient not use the same term as the doctor, but he/she may not express his/her real problem adequately. Furthermore, from a technical point of view, the corpus is generally spread across several documents, with unstandardized formats, loose structures, and highly diversified words used by patients from various backgrounds (Deleger and Zweigenbaum, 2008; Deleger, 2009).

The applicability of NLP tools is also challenging in the context of present-day psychiatric research. Available NLP tools are particularly sensitive to two aspects. The first is the ability of NLP to reduce the complexity of unstructured texts. The second is its ability to grasp the interrelations between words or concepts in a relevant way. Because of these limitations, NLP systems can at the moment only provide very basic analyses. And this is particularly true in psychiatry

where patients are often described in terms of emotions or personality by subtle notions. In addition, NLP tools are almost exclusively designed to explore texts in English. For other languages, tools either do not exist or can be used only for very basic analyses. The limitations of NLP approaches need to be identified but it is possible that the increasingly rapid advances in NLP will address these challenges. Currently, a large amount of research dedicated to web search engines is ongoing and it could have a direct impact on techniques that can be used in medical research. Finally, large textual datasets are available and cannot be analysed with other tools than NLP. Patients' messages shared on the Internet, or medical files stored on computers are sources of information that cannot be ignored. NLP approaches, even if they have obvious limitations at the moment, are likely to become essential tools for psychiatric research.

In conclusion, at a time when there is a debate on the relative merits of qualitative and quantitative methods in psychiatric research (Falissard *et al.*, 2013), TM offers an original approach. Exploratory by nature, processing free speech or texts obtained from patients or physicians, it is in many ways close to qualitative methods. It however relies heavily on sophisticated statistical and algorithmic routines, the user has a limited impact on the analysis itself, and in these aspects it is close to quantitative methods. But, above all, TM is at the moment the only family of tools that is liable to cope with the huge amount of textual data that is accumulating every day in the field of mental health, whether from medical files, patient forums or social networks. No doubt, for this simple reason, it is set to occupy an important place in the methodological landscape of psychiatric research.

References

Agarwal S., Yu H., Kohane I. (2011) BioNOT: a searchable database of biomedical negated sentences. *BMC Bioinformatics*, **12**, 420, DOI: 10.1186/1471-2105-12-420

Ananiadou S.M., McNaught J. (2006) Text Mining for Biology and Biomedicine, Boston, MA, Artech House.

Ananiadou S., Pyysalo S., Tsujii J., Kell D.B. (2010) Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, **28**(7), 381–390, DOI: 10.1016/j.tibtech.2010.04.005

Bernardi L., Tuzzi A. (2011) Analyzing written communication in AAC contexts: a statistical perspective. *Augmentative and Alternative Communication*, **27**(3), 183–194, DOI: 10.3109/07434618.2011.610353

Cameron D., Smith G.A., Daniulaityte R., Sheth A.P., Dave D., Chen L., Anand G., Carlson R., Watkins K.Z., Falck R. (2013) PREDOSE: a semantic web platform for drug abuse epidemiology using social media. *Journal of Biomedical Informatics*, **46**(6), 985–997, DOI: 10.1016/j.jbi.2013.07.007

Cheng D., Knox C., Young N., Stothard P., Damaraju S., Wishart D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, **36**(Web Server issue), W399–W405, DOI: 10.1093/nar/gkn296

Cohen K.B., Hunter L. (2008) Getting started in text mining. *PLoS Computational Biology*, **4**(1), e20, DOI: 10.1371/journal.pcbi.0040020

Cohen T., Blatter B., Patel V. (2008) Simulating expert clinical comprehension: adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics*, **41**(6), 1070–1087, DOI: 10.1016/j.jbi.2008.03.008

Cunningham H., Tablan V., Roberts A., Bontcheva K. (2013) Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, **9**(2), e1002854, DOI: 10.1371/journal.pcbi.1002854

Deleger L. (2009) Exploitation de corpus parallèles et comparables pour la détectionde correspondances lexicales: application au domaine médical, PhD thesis, Pierre et Marie Curie University, Paris.

Deleger L., Zweigenbaum P. (2008) Paraphrase acquisition from comparable medical corpora of specialized and lay texts. *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*, pp. 146–150, Bethesda, MD: AMIA.

Dias A.M., Mansur C.G., Myczkowski M., Marcolin M. (2011) Whole field tendencies in transcranial magnetic stimulation: A systematic review with data and text mining. *Asian Journal of Psychiatry*, **4**(2), 107–112, DOI: 10.1016/j.ajp.2011.03.003

Eriksson R., Jensen P.B., Frankild S., Jensen L.J., Brunak S. (2013) Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association*, **20**(5), 947–953, DOI: 10.1136/amiajnl-2013-001708

Evans J.A., Rzhetsky A. (2011) Advancing science through mining libraries, ontologies, and communities. *Journal of Biological Chemistry*, **286**(27), 23659–23666, DOI: 10.1074/jbc.R110.176370

Falissard B., Revah A., Yang S., Fagot-Largeault A. (2013) The place of words and numbers in psychiatric research. *Philosophy, Ethics, and Humanities in Medicine*, **8**, 18, DOI: 10.1186/1747-5341-8-18

Gara M.A., Vega W.A., Lesser I., Escamilla M., Lawson W.B., Wilson D.R., Fleck D.E., Strakowski S.M. (2010) The role of complex emotions in inconsistent diagnoses of schizophrenia. *Journal of Nervous and Mental Disease*, **198**(9), 609–613, DOI: 10.1097/NMD.0b013e3181e9dca9

Garfield D.A., Rapp C., Evens M. (1992) Natural language processing in psychiatry. Artificial intelligence technology and psychopathology. *Journal of Nervous and Mental Disease*, **180**(4), 227–237, DOI: 0022-3018/92/1804-0227$03.00/0

Girirajan S., Truong H.T., Blanchard C.L., Elsea S.H. (2009) A functional network module for Smith-Magenis syndrome. *Clinical Genetics*, **75**(4), 364–374, DOI: 10.1111/j.1399-0004.2008.01135.x

Gong L., Yan Y., Xie J., Liu H., Sun X. (2012) Prediction of autism susceptibility genes based on association rules. *Journal of Neuroscience Research*, **90**(6), 1119–1125, DOI: 10.1002/jnr.23015

Int. J. Methods Psychiatr. Res. 25(2): 86–100 (2016). DOI: 10.1002/mpr

Han C., Yoo S., Choi J. (2011) Evaluation of co-occurring terms in clinical documents using latent semantic indexing. *Healthcare Informatics Research*, **17**(1), 24–28, DOI: 10.4258/hir.2011.17.1.24

He Q., Veldkamp B.P., de Vries T. (2012) Screening for posttraumatic stress disorder using verbal features in self narratives: a text mining approach. *Psychiatry Research*, **198**(3), 441–447, DOI: 10.1016/j.psychres.2012.01.032

Jorge-Botana G., Olmos R., Leon J.A. (2009) Using latent semantic analysis and the predication algorithm to improve extraction of meanings from a diagnostic corpus. *Spanish Journal of Psychology*, **12**(2), 424–440.

Keski-Rahkonen A., Tozzi F. (2005) The process of recovery in eating disorder sufferers' own words: an Internet-based study. *International Journal of Eating Disorders*, **37**(Supplement), S80–S86; discussion S87–S89, DOI: 10.1002/eat.20123

Korhonen A., Seaghdha D.O., Silins I., Sun L., Hogberg J., Stenius U. (2012) Text mining for literature review and knowledge discovery in cancer risk assessment and research. *PLoS One*, **7**(4), e33427, DOI: 10.1371/journal.pone.0033427

Liu Q.Y., Sooknanan R.R., Malek L.T., Ribecco-Lutkiewicz M., Lei J.X., Shen H., Lach B., Walker P.R., Martin J., Sikorska M. (2006) Novel subtractive transcription-based amplification of mRNA (STAR) method and its application in search of rare and differentially expressed genes in AD brains. *BMC Genomics*, **7**, 286, DOI: 10.1186/1471-2164-7-286

Luo S.X., Peterson B.S., Gerber A.J. (2012) *Semantic Mapping of Social Language*: *Comparing Normal Subjects to Patients With Autism Spectrum Disorders*, Society of Biological Psychiatry 67th Annual Scientific Convention and Program, Philadelphia, PA, Society of Biological Psychiatry.

Luther S., Berndt D., Finch D., Richardson M., Hickling E., Hickam D. (2011) Using statistical text mining to supplement the development of an ontology. *Journal of Biomedical Informatics*, **44**(Supplement 1), S86–S93, DOI: 10.1016/j.jbi.2011.11.001

Malhotra A., Younesi E., Gundel M., Muller B., Heneka M.T., Hofmann-Apitius M. (2014) ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's & Dementia*, **10**(2), 238–246, DOI: 10.1016/j.jalz.2013.02.009

Meystre S.M., Savova G.K., Kipper-Schuler K.C., Hurdle J.F. (2008) Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, **47**(Suppl 1), 128–144.

Miner G., Elder J., Fast A., Hill T., Nisbet R., Delen D. (2012) Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, London: Academic Press.

Moher D., Liberati A., Tetzlaff J., Altman D.G. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, **62**(10), 1006–1012, DOI: 10.1016/j.jclinepi.2009.06.005

Neuman Y., Cohen Y., Assaf D., Kedma G. (2012) Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine*, **56**(1), 19–25, DOI: 10.1016/j.artmed.2012.06.001

Perlis R.H., Iosifescu D.V., Castro V.M., Murphy S.N., Gainer V.S., Minnier J., Cai T., Goryachev S., Zeng Q., Gallagher P.J., Fava M., Weilburg J.B., Churchill S.E., Kohane I.S., Smoller J.W. (2012) Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological Medicine*, **42**(1), 41–50, DOI: 10.1017/S0033291711000997

Pestian J., Nasrallah H., Matykiewicz P., Bennett A., Leenaars A.. (2010) Suicide note classification using natural language processing: a content analysis. *Biomedical Informatics Insights*, **2010**(3), 19–28.

Piolat A., Bannour R. (2009) An example of text analysis software (EMOTAIX-Tropes) use: the influence of anxiety on expressive writing. *Current Psychology Letters*, **25**(2), 2–21.

Ranney M.L., Choo E.K., Cunningham R.M., Spirito A., Thorsen M., Mello M.J., Morrow K. (2014) Acceptability, language, and structure of text message-based behavioral interventions for high-risk adolescent females: a qualitative study. *Journal of Adolescent Health*, **55**(1), 33–40, DOI: 10.1016/j.jadohealth.2013.12.017

Roque F.S., Jensen P.B., Schmock H., Dalgaard M., Andreatta M., Hansen T., Soeby S., Bredkjaer S., Juul A., Werge T., Jensen L.J., Brunak S. (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology*, **7**(8), e1002141, DOI: 10.1371/journal.pcbi.1002141

Rzhetsky A., Seringhaus M., Gerstein M.B. (2009) Getting started in text mining: part two. *PLoS Computational Biology*, **5**(7), e1000411, DOI: 10.1371/journal.pcbi.1000411

Sager N., Friedman C., Lyman M.S. (1987a) Computer Processing of Narrative Information, Boston, MA: Addison-Wesley.

Sager N., Friedman C., Lyman M.S. (1987b) Information Formatting of Medical Literature, Boston, MA: Addison-Wesley.

Sarkar I.N. (2012) A vector space model approach to identify genetically related diseases. *Journal of the American Medical Informatics Association*, **19**(2), 249–254, DOI: 10.1136/amiajnl-2011-000480

Shang Y., Li Y., Lin H., Yang Z. (2011) Enhancing biomedical text summarization using semantic relation extraction. *PLoS One*, **6**(8), e23862, DOI: 10.1371/journal.pone.0023862

Shiner B., D'Avolio L.W., Nguyen T.M., Zayed M.H., Watts B.V., Fiore L. (2012) Automated classification of psychotherapy note text: implications for quality assessment in PTSD care. *Journal of Evaluation in Clinical Practice*, **18**(3), 698–701, DOI: 10.1111/j.1365-2753.2011.01634.x

Sohn S., Kocher J.P., Chute C.G., Savova G.K. (2011) Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, **18**(Supplement 1), i144–i149, DOI: 10.1136/amiajnl-2011-000351

Sorensen A.A. (2009) Alzheimer's disease research: scientific productivity and impact of the top 100 investigators in the field. *Journal of Alzheimer's Disease*, **16**(3), 451–465, DOI: 10.3233/JAD-2009-1046

Tu S.W., Tennakoon L., O'Connor M., Shankar R., Das A. (2008) Using an integrated ontology and information model for querying and reasoning about phenotypes: the case of autism. *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*, pp. 727–731, Bethesda, MD: AMIA.

Veale D., Poussin G., Benes F., Pepin J.L., Levy P. (2002) Identification of quality of life concerns of patients with obstructive sleep apnoea at the time of initiation of continuous positive airway pressure: a discourse analysis. *Quality of Life Research*, **11**(4), 389–399.

Wallace B.C., Small K., Brodley C.E., Lau J., Schmid C.H., Bertram L., Lill C.M., Cohen J.T., Trikalinos T.A. (2012) Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in Medicine*, **14**(7), 663–669, DOI: 10.1038/gim.2012.7

Wang X., Hripcsak G., Markatou M., Friedman C. (2009) Active computerized pharmacovigilance using natural language processing, statistics, and

electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, **16**(3), 328–337, DOI: 10.1197/jamia.M3028

Wu J.L., Yu L.C., Chang P.C. (2012) Detecting causality from online psychiatric texts using inter-sentential language patterns. *BMC Medical Informatics and Decision Making*, **12**, 72, DOI: 10.1186/1472-6947-12-72

Wu C.Y., Chang C.K., Robson D., Jackson R., Chen S.J., Hayes R.D., Stewart R. (2013) Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PLoS One*, **8**(9), e74262, DOI: 10.1371/journal.pone.0074262

Yang S., Kadouri A., Revah-Levy A., Mulvey E.P., Falissard B. (2009) Doing time: a qualitative study of long-term incarceration and the impact of mental illness. *International Journal of Law and Psychiatry*, **32**(5), 294–303, DOI: 10.1016/j.ijlp.2009.06.003

Yu L.C., Wu C.H. (2007) Psychiatric consultation record retrieval using scenario-based representation and multilevel mixture model *IEEE Transactions on Information Technology in Biomedicine*, **11**(4), 415–427.

Yu S., Van Vooren S., Tranchevent L.C., De Moor B., Moreau Y. (2008) Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, **24**(16), i119–i125, DOI: 10.1093/bioinformatics/btn291

Yu L.-C., Wu C-H., Jang F.-L. (2009) Psychiatric document retrieval using a discourse-aware model. *Artificial Intelligence*, **173**(7–8), 817–829.

Yu L.-C., Chan C.-L., Lin C.-C., Lin I.C. (2011) Mining association language patterns using a distributional semantic model for negative life event classification. *Journal of Biomedical Informatics*, **44**(4), 509–518.

Zhang J., Dong N., Delprino R., Zhou L. (2009) Psychological strains found from in-depth interviews with 105 Chinese rural youth suicides. *Archives of Suicide Research*, **13**(2), 185–194, DOI: 10.1080/13811110902835155

Zhu F., Patumcharoenpol P., Zhang C., Yang Y., Chan J., Meechai A., Vongsangnak W., Shen B. (2013) Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, **46**(2), 200–211, DOI: 10.1016/j.jbi.2012.10.007

Zweigenbaum P., Demner-Fushman D., Yu H., Cohen K.B. (2007) Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, **8**(5), 358–375, DOI: 10.1093/bib/bbm045