Introduction

Cancer has become one of the most important threats to human life and health, and its incidence and mortality rates have been increasing steadily since the beginning of the 20th century. People diagnosed with cancer and died of the disease reached 10.1 million and 6.2 million respectively in 2000 (Dong et al., 2002) and these figures increased to 10.86 and 6.73 million only two years later (Stewart et al., 2003). It is estimated that, by 2020, people living with cancer will add up to 30 million and annual new cases and deaths of the disease will rise to 15 million and 10 million (Parkin et al., 2005). Cancer causes not only enormous physical damages but also severe psychological and social sufferings. Numerous researches have documented close association between symptoms of depression and anxiety and elevated risk and mortality of almost all cancers (Su et al., 2004; Yu et al., 2006; Pandey et al., 2006; Song, 2011; Kye et al., 2012) Surveys in China revealed that scores about depression, anxiety, obsessive compulsion, hostility, bigotry, interpersonal impairments and other disorders were all much higher in cancer patients than in norm controls (Akechi et al., 2008; Jadoon et al., 2010; Chintamani et al., 2011). Previous Japanese studies also reported 15–40% prevalence rate of psychological distress in cancer patients (Gregurek et al., 2010). Rudolf Gregurek and colleagues’ study found that one third of cancer patients experienced distress which requires evaluation and treatment, and psychiatric disorders e.g., depression, anxiety disorders and adjustment disorders, were most common (Mystakidou et al., 2005). Other studies showed that anxiety and depression in cancer patients may be caused by various reasons including psychological reaction caused by diagnosis of cancer, long duration of treatment, side effects of treatment, repeated hospitalizations, disruption in life and diminished quality of life (Jacobsen et al., 2008).

Giving the above evidences and others, psychological and behavioral intervention against cancer (PBIC) is gaining recognition and publications in this regard have been growing rapidly (Glas et al., 2001; Xie et al., 2006; Zheng et al., 2007; Barbara et al., 2008; Jiao et al., 2009; Hajian S et al., 2011). This in turn makes it more and more challenging for researchers to keep abreast with research developments, especially derive meaningful trends, patterns and linkages between issues hidden
under huge and fast growing literature. As government and non-government agencies are developing and maintaining more and more efficient and sophisticated scientific databases, literature retrieval has become very convenient and by entering a few key words, researchers can get instant access to almost all published articles of interest like PBIC. However, understanding or deriving useful information from these publications still depends primarily on manual reading, a time-consuming and painstaking task. Meanwhile, with the explosion of computer and internet technologies, easily available and applicable software and tools for text mining and network analysis are burgeoning (Oda et al., 2008; Yang et al., 2009). These provide great potential for leveraging more productive and efficient use of existing literature (Ananiadou et al., 2005; Zweigenbaum et al., 2005; Altman et al., 2007; Jelier et al., 2008; Frijters et al., 2010; U.S. National Library of Medicine, 2011). Yet this potential is far from well explored. Although the majority of researchers are used to take the advantage of scientific databases in retrieving published papers, few of them are commonly using available text mining methods, for example, to inform their conventional literature review. This study tries to perform a systematic review of literature on PBIC using terms analysis in a hope of identifying potential trends, patterns and links and deriving lessons applying text mining techniques to inform traditional literature reviews in this specific area.

Materials and Methods

Literature retrieval

The study used PubMed, one of the most popular literature databases being developed and maintained by US National Library of Medicine, as the data source. Literature retrieval utilized PubMed inbuilt search engine by entering “(cancer [Title/Abstract] AND (psychological [Title/Abstract] OR behavior [Title/Abstract]) AND intervention [Title/Abstract]” as the search algorithms and setting publication time period as from January 1, 2002 to December 31, 2011. The searched result was then “send to” a “XML” file for forthcoming processing and analysis. The search was performed at May 17, 2012.

Data coding

The XML file was then translated into dichotomized records (one record for each article) representing presence or non-presence of MeSH terms and a metric consisting of numbers of times of co-occurrence between all pairs of terms identified via a mini program developed by ourselves using Microsoft Visual Studio 2008 and SQL2008. MeSH terms here refers to the Medical Subheading Tree version 2012 which comprises 54092 distinct terms (Liu et al., 2006).

Data analysis and visualization

Data analysis used the records generated above via Microsoft Excel2007. It consisted of counting the numbers of occurrence of MeSH terms and calculating term occurrence probability (hereafter referred to as TOP). The TOP of an individual term for a given year denotes the percentage of times of the term occurred in all articles to the total number of the articles in the same year. Data visualization included histograms (Figure 1) and terms networks (Figures 2 and Figure 3). All the histograms were produced via EXCELL2007 using the aforementioned records; while the terms networks, using UCINET (Zong et al., 2011) and the metric produced above.

Results

Occurrence of MeSH terms

A total of 997 articles were selected into this study which included 1742 MeSH terms. Put in descending frequency, these terms demonstrate a typical hyperbolic curve (Killeen, 2011; Thessen et al., 2012). It shows that high frequency terms accounted for a very small proportion of all terms occurred; while the majority of terms appeared in only a few articles. More specifically, terms occurred less than 10 times added up to 1458 (83.69% in total) while those occurred 10-100 times, 101-500 times and over 500 times accounted for 247 (14.18%), 29 (1.67%) and 6 (0.3%) respectively.

Figure 1 depicts the “finger prints” of PBIC terms for 2002 (upper half) and 2011 (lower half) in which, the vertical lines representing TOPs were plotted along the horizontal axis representing the whole set of MeSH terms arranged in the order as they appear in MeSH Trees 2012. It reveals that PBIC researches focused primarily on 5 major clusters of terms located within MO1-3000, MO9000-15000, MO16000-19000, MO39000-42000, MO44000-54000 respectively (MO here stands for Mesh Order).

Changes in term occurrence

Figure 1 also invokes comparison of the TOPs between 2002 and 2011. It shows that although there witnessed a substantial increase in total articles published from 2002 to 2011, the general pattern of the TOPs remained unchanged, i.e., the position and format of the main peaks and pitches in the TOPs of the two years resemble each other. Subtraction of the TOPs of 2011 with that of 2002 revealed that although most TOPs differed between the two years, terms with substantial (e.g., over 5%) TOP
Table 1. Terms Found with the Largest TOP Increase or Decrease: TOP2011 Minus TOP2002

<table>
<thead>
<tr>
<th>Terms</th>
<th>%</th>
<th>Terms</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survivors</td>
<td>13.7</td>
<td>etiology</td>
<td>-17.1</td>
</tr>
<tr>
<td>Young Adult</td>
<td>11.0</td>
<td>Risk Assessment</td>
<td>-7.5</td>
</tr>
<tr>
<td>Quality of Life</td>
<td>10.9</td>
<td>Women’s Health</td>
<td>-6.8</td>
</tr>
<tr>
<td>Depression</td>
<td>10.8</td>
<td>administration &amp; dosage</td>
<td>-6.7</td>
</tr>
<tr>
<td>statistics &amp; numerical data</td>
<td>10.0</td>
<td>Attitude to Health</td>
<td>-6.5</td>
</tr>
<tr>
<td>Anxiety</td>
<td>9.8</td>
<td>Time</td>
<td>-6.4</td>
</tr>
<tr>
<td>Breast</td>
<td>9.4</td>
<td>Women</td>
<td>-6.0</td>
</tr>
<tr>
<td>Breast Neoplasms</td>
<td>9.4</td>
<td>Questionnaires</td>
<td>-6.0</td>
</tr>
<tr>
<td>Exercise</td>
<td>7.9</td>
<td>Time Factors</td>
<td>-5.8</td>
</tr>
<tr>
<td>psychology</td>
<td>7.6</td>
<td>Health</td>
<td>-5.8</td>
</tr>
<tr>
<td>Cognitive Therapy</td>
<td>6.8</td>
<td>Mental Health</td>
<td>-5.3</td>
</tr>
<tr>
<td>Early Detection of Cancer</td>
<td>6.8</td>
<td>Men</td>
<td>-5.2</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>6.6</td>
<td>Oncologic Nursing</td>
<td>-5.2</td>
</tr>
<tr>
<td>rehabilitation</td>
<td>6.3</td>
<td>Adaptation, Psychological</td>
<td>-4.9</td>
</tr>
<tr>
<td>psychotherapy</td>
<td>5.7</td>
<td>Cohort Studies</td>
<td>-4.6</td>
</tr>
<tr>
<td>Prevalence</td>
<td>5.1</td>
<td>Reverse Transcriptase Polymerase Chain Reaction</td>
<td>-4.6</td>
</tr>
<tr>
<td>Animals</td>
<td>4.6</td>
<td>Cohort Studies</td>
<td>-4.6</td>
</tr>
<tr>
<td>Comorbidity</td>
<td>4.4</td>
<td>Skin</td>
<td>-4.6</td>
</tr>
<tr>
<td>Randomized Controlled Trials as Topic</td>
<td>4.4</td>
<td>Skin Neoplasms</td>
<td>-4.6</td>
</tr>
<tr>
<td>Patient Education as Topic</td>
<td>4.3</td>
<td>Medicine</td>
<td>-4.5</td>
</tr>
</tbody>
</table>

The TOP of an individual term for a given year denotes the percentage of times of the term occurred in all articles to the total number of the articles in the same year.

Figure 2. USINET- GMS Network of High Frequency Terms. The figure was drawn using the GMS (Gower metric scaling layout) of UCINET in which the relation between any pair of terms was measured in times of co-occurrence of the two terms and the threshold of times of co-occurrence for judging the presence of connections was set as 20. The figure visualizes the distance between terms and shows roughly four major cliques within the network (up triangle, square, diamond, circle) which may be labeled as therapy and outcomes, psycho-behavior intervention, methodology and population, and miscellaneous cliques respectively.

Figure 3. USINET- OSM Network of High Frequency Terms. The figure was drawn using the OSM (layout using options set in menu) of UCINET in which the relation between any pair of terms was measured in times of co-occurrence of the two terms and the threshold of times of co-occurrence for judging the presence of connections was set as 20; and diamond, up triangle, square, circle stands for therapy and outcomes, psycho-behavior intervention, methodology and population, and miscellaneous cliques respectively. It consists of some 200 interconnected terms arranged in a centripetal way, i.e., the terms with larger number of lines connected with others (e.g., humans, neoplasms, female) were put more close to the center of the figure, while those with smaller number of relations (e.g., animals, lung, sleep), were plotted nearer to the edge of the network. In addition, the position of a term in the network also reflects its closeness to its neighboring terms. Based on these rules, a series of

S

Theme relations

Figure 2 is a GMS (Gower Metric Scaling Layout) network of our PBIC terms (using the metric mentioned in the methodology section) produced via UCINET-NetDraw, in which the relation between any pair of terms was measured in times of co-occurrence of the two terms and the threshold of times of co-occurrence for judging the presence of connections was set as 20. The figure visualizes the distance between terms and shows roughly
patterns or characteristics was observed from the figure (Table 2).

**Discussion**

The findings concerning the term occurrence inform scope and content of PBIC studies. Nearly two thousand distinct terms identified from the publications imply that PBIC comprises quite a huge area of research interests or issues. The “term print” provides a rapid means for generating, from a vast literature, an overview about what were the hot and cold topic regions among PBIC research communities. It may also be used as a visual aid for exploring new PBIC applications or research issues. A region clear of any vertical lines along the x-axis in figure 1 means it is a region either not applicable to or not yet explored for PBIC. Therefore, by checking into the MeSH terms presented by the numbers of the “clear” regions, researchers may identify novel applications or issues for PBIC. While for the hyperbolic curve of PBIC terms, the so called “small science” also applies. In other words, to the left of the curve are the few terms (e.g., humans, female, neoplasms) being highly recognized by most researchers; while to the curve are the many terms producing over 80% of scientific output.

The changes in term occurrence tell trends of PBIC researches. The large discrepancy (462 vs. 643) between the absolute number of total terms identified in 2011 and that in 2002 suggests substantial expansion in PBIC research scope. The differences in TOPs of individual terms between the two years show changes in relative focus of PBIC researches. The terms found with the largest magnitude of TOP increase or decrease (Table 1) can be classified into five main categories, i.e., terms about population and subjects (PS, e.g., male, female, adult, adolescent), intervention approaches (IA, e.g., radiotherapy, exercise, psychotherapy), disease and etiology (DE, e.g., skin neoplasms, breast neoplasms), methodology and techniques (MT, e.g., Randomized Controlled Trails, longitudinal studies), and outcomes (OC, e.g., depression, anxiety, stress). Examining the TOP increases and decreases given in table 1 by these categories, it is arguable that: a) the focus of PS-related issues was shifting from general groups (e.g., male, female) to priority groups (e.g., survivors, adolescents) and animals; b) IA was emphasizing psycho-behavior measures (e.g., exercise, psychotherapy) than traditional clinical treatment (e.g., medicine, radiotherapy); c) DE priority was moving from skin neoplasms to breast neoplasms; d) MT was preferring rigid yet cost-efficient approaches (randomized controlled trials) to sophisticated and costly ones (e.g., Reverse Transcriptase Polymerase Chain Reaction, cohort studies); and e) OC was stressing quality of life, anxiety, depression etc. instead of health and mental health.

The term networks provide intuitive means for reviewing various relations between researched issues. The GMS network (Figure 2) leads to a number of interesting inferences. The composition and structure of the yellow clique suggests that: quality of life, anxiety, depression, complications, treatment outcome and fatigue etc. were among the priority issues and they were often addressed together within an individual article; strategies used to reach these outcomes included
mainly therapy, psychotherapy, psycho-adaptation, social support, rehabilitation, exercise etc.; and comprehensive intervention (i.e., combining therapy with psychotherapy, social support etc. rather than therapy alone) had become a common strategy against cancer. The red clique indicates that: health knowledge, attitudes and practices (KAP) were another set of important and highly co-occurring topics; health education, patient education, prevention and control programs were chief measures employed to archive KAP changes; and the primary goal of these efforts were controlling cancer risk and risk factors. Similarly, the green clique suggests that: aged and middle aged were “neutral” groups for all kinds of studies (they were located at the center and had roughly equal distance to all other major cliques); adult, male, breast neoplasms were more frequently used as subjects for general methodology (including questionnaires, diagnosis, and epidemiology) studies; adolescents were the primary group for behavior intervention; and survivors and the 80 and over group were more closely linked to therapy and outcome-related studies. The black clique comprised the majority of terms clustered around human and neoplasm. This clique may not necessarily reflect meaningful links. Most terms within this clique occurred only a few times more than the threshold for inclusion and thus they were seldom linked with other terms except for the two most frequent terms human and neoplasm.

Unlike the GMS network which depicts cliques of and distances between terms, the OSM network tries to clearly visualize as many terms as possible included. Given the large number of terms an OSM network generally has, it conveys enormous information. Yet how to draw the information is challenging. We here propose two pragmatic strategies, namely concentric circle-wise (CC) and concentric arc-wise (CA) approaches. Combined with adequate background knowledge, these approaches should facilitate systematically deriving implications and clues from OSM networks. CC approach involves three steps: a) draw concentric circles using the most popular term (humans in our case) as the center and divides the network into a few (e.g., three) circled areas (see Figure 5b); b) examine, in an inward or outward order, the terms fallen into each of the circles; and c) interpret the terms within different areas differently, i.e., terms within more inward circles occurred more frequently and were more highly recognized, while those in more outward circles were more novel or irrelevant. CA approach works in similar processes: a) divide the whole network into equal arcs of an appropriate (e.g., 30°) using the most popular term as the center (see Figure 3); b) examine the network round by round in a clockwise manner; c) each round of examination involves four consecutive arcs using the middle two as “core” arcs and the remaining two as “reference” arcs; d) interpret the relations among the terms within the core arcs incorporating information from the reference arcs and background knowledge. Table 2 shows not only interesting patterns and features of PBIC studies in the past decade but also an example for identifying hidden information from complex GSM terms networks using CA approach.

Acknowledgements

This study was partly supported by China Natural Science and Technology Foundation Grant: 81172201) and Anhui Foundation for Excellent Young Scholars (Grant: 2011SQRL060). The author(s) declare that they have no competing interests.

References


