Observational Study

Medicine®

OPEN

Cancer studies based on secondary data analysis of the Taiwan's National Health Insurance Research Database

A computational text analysis and visualization study

Jui-Kun Chiang, MD, MSc[a], Chih-Wen Lin, MD[b,c], Chun-Lung Wang, MD[d], Malcolm Koo, PhD[e,f,*], Yee-Hsin Kao, MD[g,*]

Abstract

There has been a surge in the academic publication output based on secondary analyses of the data from the Taiwan's National Health Insurance claim records. It has become a challenge to comprehend such a rapid expansion of the literature. Therefore, this study aimed to explore the conceptual content of National Health Insurance Research Database-based cancer research, using the abstract of articles extracted from PubMed between 2002 and 2015. Search terms including "National Health Insurance Research Database (NHIRD) AND Taiwan," "Taiwan AND population-based," and "Taiwan AND nationwide" were used to search in PubMed with the publication date limited to between 1997 and 2015. The retrieved articles were manually screened to retain only those that were cancer-related and were based on secondary data analysis of the NHIRD. A total 589 articles were selected for subsequent text mining using the R software. Among the 589 articles, the top 5 most studied cancer types were breast (16.3%), lung (11.4%), colorectal (10.4%), liver (8.3%), and prostate (7.5%). The articles that received the highest number of citations by PubMed Central articles were cited 92 times. The top 3 most frequently occurred keywords in the abstracts of the 589 articles were cancer, patient, and risk, with 3670, 2535, and 1652 times, respectively. Analysis of key conception indicated that the most common conceptions were diabetes, survival, breast cancer, lung cancer, and colorectal cancer. In conclusion, in this study of 589 published articles on secondary data analysis of the NHIRD, indexed by PubMed between 2002 and 2015, we found that while the risk factors of cancer, treatment of cancer, and survival of cancer patients were popular research topics, end-of-life cancer care issues were less studied. Further studies should explore these areas since they are as important as treatment of the disease itself for many patients.

Abbreviations: CI = confidence interval, HR = hazard ratio, NHI = National Health Insurance, NHIRD = National Health Insurance Research Database, OR = odds ratio.

Keywords: bibliometric analysis, cancer, National Health Insurance Research Database, PubMed

1. Introduction

Despite advances in the diagnosis and treatment, cancer is still a leading cause of death worldwide. There were 14.1 million new cases of cancer and 8.2 million cancer deaths worldwide in 2012.[1] According to the World Health Organization estimates for 2011, cancer causes more deaths than coronary heart disease or stroke.[2] In addition to primary studies designed specifically for cancer research, secondary analysis of existing data collected for nonresearch purposes has been used as a cost-effective approach to complement findings from primary studies and to help explore new research hypotheses.[3]

In Taiwan, a government-run, single-payer National Health Insurance (NHI) scheme was established in 1995. The coverage rate is over 99% of Taiwan's 23 million residents. More than 20,000 medical care facilities, including hospitals, clinics, pharmacies, and medical laboratories, which represent over 93% of all healthcare facilities in Taiwan, were contracted by the NHI scheme.[4] Under the universal health coverage scheme, virtually all healthcare services, including consulting and treatment expenses for inpatient and ambulatory care, dental services, traditional Chinese medicine therapies, physical rehabilitation, and home care, are being covered. Enrollees of the scheme are issued with an integrated circuit-embedded smart card that is used to obtain medical services. For contracted healthcare organizations

Editor: Mohammad Derakhshan.

J-KC, C-WL, and C-LW have contributed equally to this work.

The authors have no funding and conflicts of interest to disclose.

[a] Division of Family Medicine, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, [b] Division of Medical Imaging, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Dalin, Chiayi, [c] School of Medicine, Tzu Chi University, Hualien, [d] Division of Pediatrics, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, [e] Department of Medical Research, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Dalin, Chiayi, Taiwan, [f] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada, [g] Division of Family Medicine, Tainan Municipal Hospital, Tainan City, Taiwan.

[*] Correspondence: Malcolm Koo, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Dalin, Chiayi, Taiwan (e-mail: m.koo@utoronto.ca)., Yee-Hsin Kao, Division of Family Medicine, Tainan Municipal Hospital, Tainan City, Taiwan (e-mail: m2200767@gmail.com).

Medicine (2017) 96:17(e6704)

Received: 10 November 2016 / Received in final form: 20 March 2017 / Accepted: 31 March 2017

http://dx.doi.org/10.1097/MD.0000000000006704

to receive reimbursement, they must submit relevant claim records to the NHI Administration (the former Bureau of National Health Insurance). Claims are then reviewed by a panel of medical experts for the type, volume, quality, and appropriateness of medical services provided under the NHI program. Diagnosis or use of medical services that are not conformed to the NHI fee schedule, drug list, clinical guidelines, and patient conditions can result in a severe penalty. Under the NHI system, certain diseases or injuries are classified as catastrophic illnesses. Patients with these illnesses can apply for a certificate, which allows them to waive outpatient and inpatient copayments. For example, the insurance of the certificate for cancer requires a diagnosis by physicians with pathological reports and a formal review by the NHI Administration.

The claim records of the NHI are established as a database, the National Health Insurance Research Database (NHIRD), available for application by eligible scientists in Taiwan for research purposes since 1998. The database consisted of the original claim records and a number of different linkable registration files. The registration files contain information on contracted medical facilities, medical personnel, and drug prescriptions. Researchers can apply for specific subject dataset such as cancer dataset or catastrophic illness dataset as well as longitudinal dataset (Longitudinal Health Insurance Database) containing a random sample of 1 million NHI enrollees.[5] The availability of such population-based database has stimulated and facilitated academic research in various scientific disciplines, especially in the area of health research.[6] The exceedingly rapid expansion of the literature based on NHIRD has made it a challenge to comprehend what has already been done, particularly in broad subject areas such as cancer research. Therefore, the aim of this study is to explore the conceptual content of NHIRD-based cancer research based on the abstract of articles extracted from PubMed. Findings from this study may be used to identify gaps in cancer research based on the NHIRD. In addition, the broader scientific community may be able to gain ideas and insights based on the existing NHIRD studies for the development of their own primary studies.

2. Materials and methods

2.1. Data source

Data were retrieved and downloaded from PubMed, a website (http://www.ncbi.nlm.nih.gov/pubmed/) that provides free access to biomedical journal citations and abstracts mainly indexed by Medline.[7,8] The service is administered by the National Center for Biotechnology Information of the United States National Library of Medicine. Search terms including “National Health Insurance Research Database AND Taiwan,” “Taiwan AND population-based,” and “Taiwan AND nationwide” were used in the search strategies. The publication date was limited to the year between 1997 and 2015. The search was conducted on July 30, 2016 and a total of 4586 articles were retrieved. The retrieved articles were manually screened by 2 authors (Y-HK and MK) to eliminate articles that were not based on data from the Taiwan’s NHIRD (1008 articles excluded and 3578 articles remained) and on the topic of cancer (2989 articles eliminated). The resulting 589 articles were included in subsequent analyses.

2.2. Text mining and data analysis

Text mining was performed using the R 3.0.2 software (R Foundation for Statistical Computing, Vienna, Austria). Four packages for R (https://cran.r-project.org/web/packages/), including RISmed for extracting bibliographic content from PubMed, SnowballC for collapsing words to a common root to aid comparison of vocabulary, tm for text mining, and rentrez for processing the results of PubMed searches were used. Descriptive analyses were conducted to calculate the frequencies of published articles for different journals and the origin of cancer sites among the studies. Citation frequency by PubMed Central articles was obtained on July 30, 2016. We also calculated the citation frequency up to 2 years of publication for the 2 top-ranking articles. The date of publication was defined as either the publication date or the Epub date, whichever was earlier. In addition, we also reported the frequency of author self-citation with self-citation defined as exists when the citing and the cited papers have at least 1 author in common. Furthermore, word cloud diagrams were generated to visualize the original word frequencies among the abstracts from the 589 articles. The co-occurrence of keywords (keyword association) in the text of the articles’ abstracts was evaluated using correlation analysis. Two keywords occurred together in the same abstract indicate that they are associated, and the strength of their association was quantified with Pearson correlation coefficients. The top 15 most frequently occurring keywords (primary keywords) in the abstract were analyzed for the correlation with other keywords (secondary keywords) in the abstract. Next, individual keywords with similar concepts were combined into key conceptions to simplify subsequent visualization using network plots.

The study protocol was reviewed and approved by the institutional review board of Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Taiwan (No. B10501012).

3. Results

Of the 3578 publications based on the data from the NHIRD publishing between 1997 and 2015, 589 (16.5%) were on the topic of cancer. The earliest article on cancer based on the NHIRD appeared in 2002 and the number of articles grew from 2 in 2002 to 160 in 2015 (Fig. 1). Table 1 lists the journals that published more than 4 studies from NHIRD between the years 2002 and 2015. The top 3 journals with most NHIRD cancer articles were PLoS ONE, Medicine, and BMC Cancer, with 64, 37, and 19 articles, respectively. The 5-year impact factors of the

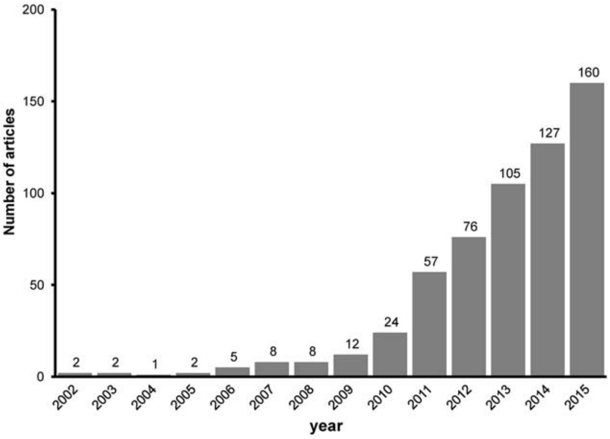

Figure 1. Number of PubMed cancer articles based on the Taiwan’s National Health Insurance Research Database from 2002 to 2015.

Table 1

Journals that published more than 4 cancer articles based on the National Health Insurance Research Database between 2002 and 2015.

Title (5-year impact factor*)	Frequency
PLoS ONE (3.535)	64
Medicine (3.195)	37
BMC Cancer (3.642)	19
British Journal of Cancer (5.617)	11
Oncologist (4.930)	10
International Journal of Cancer (5.624)	9
QJM (2.634)	8
Japanese Journal of Clinical Oncology (2.010)	8
Oral Oncology (3.752)	7
Evidence-based Complementary and Alternative Medicine (2.140)	6
Cancer (5.434)	6
Journal of the Chinese Medical Association (0.911)	6
Annals of Surgical Oncology (4.239)	6
Computer Methods and Programs in Biomedicine (2.051)	6
Diabetes Care (9.015)	5
Psycho-oncology (3.786)	5
Cancer Epidemiology, Biomarkers & Prevention (4.251)	5
Supportive Care in Cancer (2.689)	5

* Five-year impact factors were based on the 2015 Journal Citation Reports.

journals ranged from 0.91 to 9.02, with a median of 3.70. The top 5 most studied cancer types were breast (16.3%), lung (11.4%), colorectal (10.4%), liver (8.3%), and prostate (7.5%) (Table 2).

3.1. Citation frequency by PubMed central articles

Of the 589 studies, the one that received the highest number of citations by PubMed Central articles (92 times as of July 30, 2016) (37 times within 2 years of publication [i.e., up to November 14, 2014], of which 1 was author self-citations) was a study aimed to investigate the association between nucleoside analog use and risk of tumor recurrence in patients with hepatitis B virus-related hepatocellular carcinoma after curative surgery. The study, published in the Journal of the American Medical Association in 2012, used a cohort design and Cox regression analysis to calculate hazard ratios (HRs) of 518 patients treated with nucleoside analogs compared with 4051 patients without the treatment. The results showed that nucleoside analog use was independently associated with a reduced risk of hepatocellular carcinoma recurrence (HR=0.67, $P<0.001$).[9]

Table 2

Cancer types in the articles based on the National Health Insurance Research Database between 2002 and 2015 (N=589).

Cancer type	Frequency (%)
Breast cancer	96 (16.3)
Lung cancer	67 (11.4)
Colorectal cancer	61 (10.4)
Liver cancer	49 (8.3)
Prostate cancer	44 (7.5)
Cervical cancer	35 (5.9)
Lymphoma	27 (4.6)
Gastric cancer	25 (4.2)
Head and neck cancer	24 (4.1)
Thyroid cancer	22 (3.7)
Renal and urinary cancer	21 (3.6)
Leukemia	18 (3.1)
Esophageal cancer	18 (3.1)
Ovarian cancer	15 (2.5)
Endometrial cancer	11 (1.9)
Pancreatic cancer	8 (1.4)

Cancers below 1% occurrence in the articles are not listed.

The study that received the second highest number of citations by PubMed Central articles (43 times as of July 30, 2016) (8 times within 2 years of publication [i.e., up to August 5, 2011], of which 2 were author self-citations) was also published in 2009 by the same group of investigators as the above study. The cohort study, published in Gastroenterology, found that early *Helicobacter pylori* eradication was associated with a low gastric cancer risk (HR=0.77) in 80,255 patients with peptic ulcer diseases.[10]

A few studies focused on medications use had received a relatively high number of citations. A 2010 study, published in the Journal of the National Cancer Institute, reported that the consumption of aristolochic acid-containing Chinese herbal products (e.g., Mu Tong) was associated with an increased risk of cancer of the urinary tract in a dose-dependent manner that is independent of arsenic exposure. This study had been cited 36 times.[11] Another study published in 2013, which had been cited 24 times, found that statin use was associated with a reduced risk of hepatocellular carcinoma among patients with chronic hepatitis C virus infection.[12] In addition, a 2011 study based on a case–control design reported that statins might reduce the risk of liver cancer (adjusted odds ratio [OR]=0.62, 95% confidence interval [95% CI]=0.42–0.91). This study had been cited 22 times.[13]

Furthermore, we noted that NHIRD cancer studies related to diabetes received a relatively high number of citations. A secondary cohort study of 472,979 adult patients with type 2 diabetes suggested that diabetes was associated with an increased cancer risk. This study, published in 2014, had been cited 8 times by PubMed Central articles.[14] On the other hand, another secondary cohort study published in 2012 found that patients with diabetes were not at increased risk for the development of lung cancer, but the use of antidiabetes drugs could decrease the risk by up to 45%. This study had been cited 30 times.[15] A secondary case–control study published in 2012 did not find a significant association between pioglitazone and bladder cancer in 54,928 patients with type 2 diabetes (adjusted HR=1.31, 95% CI=0.66–2.58). This study had been cited 27 times.[16] Another secondary case–control study on 606,583 type 2 diabetic patients published in 2012 found that the use of pioglitazone (OR=0.73, 95% CI=0.65–0.81) and rosiglitazone (OR=0.83, 95% CI=0.72–0.95) was associated with a decreased liver cancer incidence in diabetic patients. This study had been cited 25 times.[17] Moreover, a secondary cohort study designed to examine cancer incidence associated with the use of insulin glargine versus intermediate/long-acting human insulin showed that insulin glargine use did not increase the risk of overall cancer incidence, but it was positively associated with both pancreatic cancer (adjusted HR=2.15, 95% CI=1.01–4.59) and prostate cancer (adjusted HR=2.42, 95% CI=1.50–8.40) in men. This study, published in 2011, had been cited 21 times.[18]

3.2. Word cloud

The word frequencies in the abstracts among 589 articles were visualized as a word cloud with a larger word size represents a higher frequency of appearance among the articles (Fig. 2). The top 3 words with the highest frequency both over different

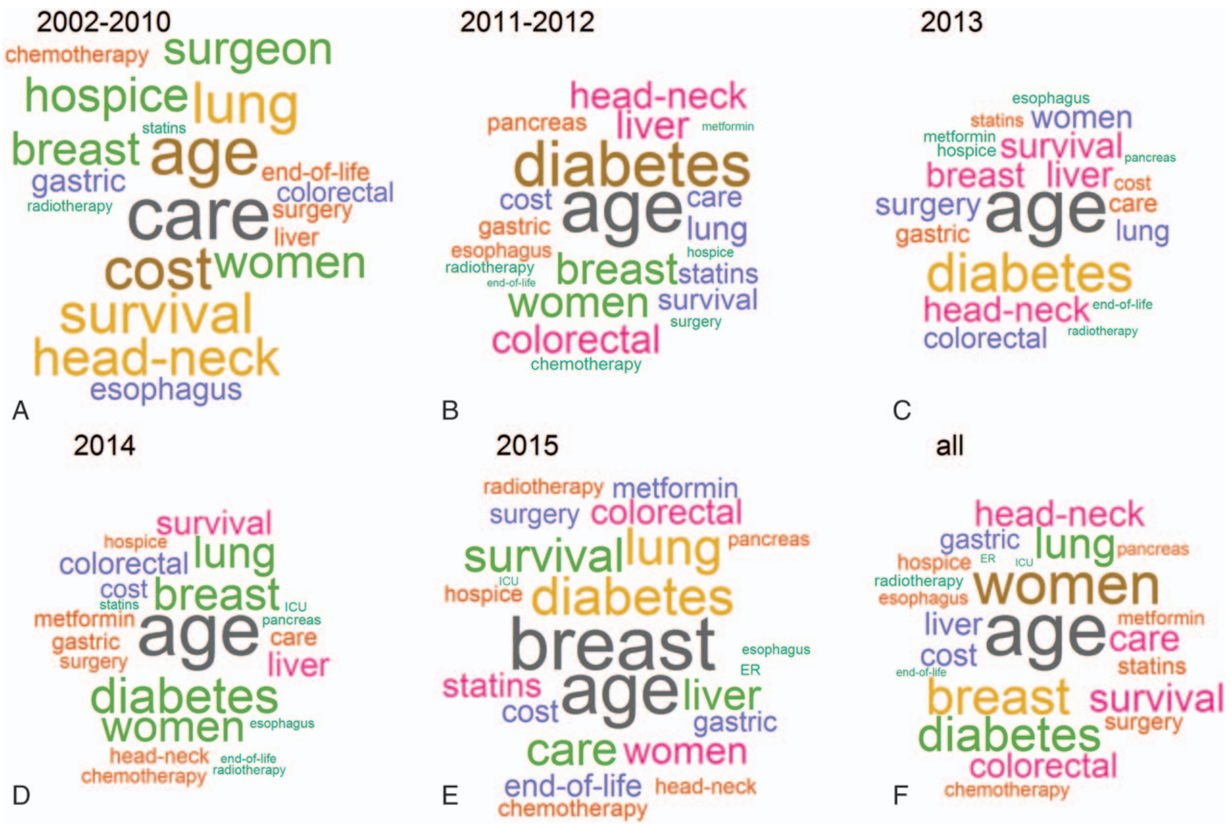

Figure 2. Plots of word cloud of keywords from the abstracts of 589 cancer articles, by different publication years. ER = emergency room, ICU = intensive care unit.

periods and over the entire period of 2002 to 2015 were cancer, patients, and risk. These 3 words were suppressed in the display of the word cloud to allow a better visualization of the remaining words. Words ranked the fourth to the sixth in frequency were care, survival, and lung in 2002 to 2010; age, breast, and women in 2011 to 2012; age, diabetes, and breast in 2013; age, breast, and women in 2014; and breast, age, and care in 2015. Over the entire period between 2002 and 2015, age, breast, and women were the words ranked the fourth to the sixth in frequency.

3.3. Analysis of keyword association

Table 3 shows the results of the analysis of keyword association for the top 15 most frequently occurred keywords in the articles' abstracts. The top 3 most frequently occurred keywords in the abstracts of the 589 articles were cancer, patient, and risk with 3670, 2535, and 1652 times, respectively. The remaining 12 keywords had a frequency of occurrence ranged from 692 times in "age" to 131 times in "hospice." Each of the primary keywords was evaluated for their correlations with other keywords in the abstract (secondary keywords). Overall, the correlation coefficients ranged from the highest at 0.68 for "neck" and "head" to the lowest at 0.12 for "cancer" and "cervix."

3.4. Analysis of key conception

A list of key conceptions was created from the keywords from the 589 articles (Table 4). To facilitate the comprehension of the subsequent plots of key conceptions, only those conceptions (nodes in Fig. 3A–E) with at least 4 connecting lines with other conceptions were shown in Table 4 and Fig. 3A–E. The most common conception was diabetes, which appeared in 20% of the articles. Next, the conception of survival appeared in 19.2% of the articles. Breast cancer, lung cancer, and colorectal cancer were the next 3 most common conceptions with an appearance of over 10% in the articles. Figure 3A–E shows the network plots for the key conceptions listed in Table 4. The data were displayed separately in 5 periods, (A) 2002 to 2010, (B) 2011 to 2012, (C) 2013, (D) 2014, and (E) 2015, to provide a clearer view of the associations among the different conceptions. For the years 2002 to 2010 (64 articles), the associations among key conceptions were the consequences of cancer (particularly breast cancer), such as the issue of patient's survival, healthcare costs, chemotherapy, hospice, and palliative care. For the years 2011 to 2012 (133 articles), the associations were about the risk factors of cancer (particularly breast cancer, prostate cancer, hepatocellular carcinoma, and lung cancer), such as diabetes, hypertension, dyslipidemia, stroke, infarction, and medications for diabetes, hypertension, and statin use. For the year 2013 (105 articles), the associations were also related to the risk factors of cancer, such as diabetes, hypertension, dyslipidemia, and medications for diabetes, hypertension, and statin use. For the year 2014 (127 articles), the key conceptions were related to the risk factors of cancer (particularly lung cancer and breast cancer), such as diabetes, hypertension, and stroke; cancer treatments, such as surgery, chemotherapy, and radiotherapy; and issue of patient's survival. For the year 2015 (160 articles), the associations were on the risk factors for a wider range of cancers (breast cancer, hepatocellular carcinoma, lung cancer, colorectal cancer,

Table 3

Analysis of keyword association of the corpus for the top 15 most frequently occurring keywords in the abstract.

Primary keyword	Frequency	Top 10 secondary keywords with highest correlation with the primary keyword (correlation coefficient)
1. Cancer	3670	Risk (0.28), breast (0.27), liver (0.20), rosiglitazone (0.19), esophagus (0.17), lung (0.17), endometrial (0.15), colorectal (0.14), age (0.13), cervix (0.12)
2. Patient	2535	Very high volume (0.42), reconstructive (0.39), excision (0.36), tumour (0.33), hierarchical (0.31), hospital (0.31), surgeons (0.31), length (0.29), operation (0.29), resection (0.27)
3. Risk	1652	Developing (0.32), ratio (0.30), cohort (0.26), hazard (0.24), patients (0.20), vagina/vulva (0.20), hematologic (0.18), insurance (0.18), cancers (0.17), lifestyle (0.17)
4. Age	692	Sex (0.30), incidence (0.24), schizophrenic (0.22), ratio (0.19), gender (0.17), nondiabetes (0.17), progesterone (0.17), men (0.16), diabetologists (0.15), retinoblastoma (0.15)
5. Breast	455	Gender (0.29), tamoxifen (0.27), formula (0.24), civil (0.23), menstrual (0.23), nurse (0.23), screening finding (0.23), self-examination (0.23), self-report (0.23), antibody (0.20)
6. Women	526	Practitioner (0.46), racial (0.46), disability (0.45), gender (0.38), disparity (0.32), routine (0.28), preventive (0.27), adenomyosis (0.23), clinics (0.23), diabetes (0.23)
7. Survival	329	Improved (0.29), prognostic (0.29), poorer (0.25), distant (0.23), stage (0.23), outcome (0.22), TNM (0.21), radiotherapy (0.20), treatment (0.20), metastases (0.19)
8. Lung	364	Pulmonary (0.43), adenocarcinoma (0.39), EGFR (0.34), COPD (0.33), squamous (0.28), obstructive (0.27), asthma (0.26), neuroendocrine (0.26), nonsmall (0.23), epidermal (0.22)
9. Diabetes	418	Type (0.56), nondiabetes (0.45), duration (0.32), nephropathy (0.32), antidiabetes (0.31), dyslipidemia (0.24), insulin (0.24), pancreatic (0.23), metabolic (0.22), age (0.21)
10. Care	292	EOL (0.59), aggressive (0.47), intubation (0.46), resuscitation (0.44), oncologists (0.34), quality (0.34), intensive care unit (0.33), hospital (0.32), home (0.30), emergency (0.29)
11. Liver	237	Cirrhosis (0.58), age (0.47), fatty (0.47), metastasis (0.36), hepatitis (0.34), nonalcoholic (0.32), operating (0.32), virus (0.31), biomedicine (0.30), farmers (0.30)
12. Colorectal	269	Cancer research (0.38), TZD (0.32), antineoplastic (0.30), faecal (0.30), NSAID (0.30), PPARγ (0.30), screenees (0.30), duration dependent (0.27), hypoglycaemic (0.27), adenoma (0.26)
13. Neck	288	Head (0.68), artery (0.63), coronary (0.55), cataract (0.51), diabetes (0.50), radiation-induced (0.50), lifestyle (0.42), mouth (0.22), chemo-radiotherapy (0.20), nutritional (0.20)
14. Gastric	190	Ulcer (0.45), pylori (0.41), duodenal (0.38), helicobacter (0.37), peptic (0.37), prevent (0.32), cardia (0.27), receptors (0.27), serotonin reuptake (0.27), year case (0.27)
15. Hospice	131	Died (0.44), noncancer (0.44), referrals (0.43), home (0.34), admissions (0.30), utilization (0.30), family (0.28), EOL (0.27), inpatient (0.27), home-based (0.24)

COPD = chronic obstructive pulmonary disease, EGFR = estimated glomerular filtration rate, EOL = end-of-life, NSAID = nonsteroidal anti-inflammatory drugs, PPARγ = peroxisome proliferator-activated receptor-γ, TNM = tumor, lymph nodes, and metastasis, TZD = thiazolidinediones.

lymphoma, and gastric cancer), such as diabetes, hypertension, dyslipidemia, and stroke. The issues of cancer treatments, such as surgery, chemotherapy, and radiotherapy and the issue of patient's survival and cancer-related healthcare costs also received a large number of associations.

4. Discussion

The present study is the first to conduct a computational text analysis and visualization of articles indexed by PubMed on cancer research that were based on secondary data analyses of the Taiwan's NHIRD. We used various approaches, including word cloud, tokens association, and conceptions network, to visualize the content of 589 articles in NHIRD published between 2002 and 2015.

We found that breast cancer, lung cancer, colorectal cancer, liver cancer, and prostate cancer were the top 5 most studied cancers based on the NHIRD data. This is not surprising because these are the most common cancer type in Taiwan. The 5 highest incidences of invasive cancers in 2013 were of the female breast, colorectal, liver, lung, and prostate.[19] The large number of cases in these cancers provides sufficient sample sizes for exploring various analyses of their risk factors, survival, and treatment. Conversely, less common cancer types do not have enough cases for investigating their associations with other disorders except when the latter have a high prevalence, such as diabetes, hypertension, and hyperlipidemia.

In the word cloud visualization, we found that survival of patients, hospice care, and end-of-life care were gradually increased from 2012 to 2015. Treatments of cancer, such as surgery, chemotherapy, and radiotherapy, were less frequently appeared words in most of the study periods, but their appearance was mildly increased from 2014 to 2015.

In the visualization of the associations of key conceptions, the most frequent associations were generally related to cancer risk factors, followed by survival, therapy-related for cancer, and end-of-life care. The associations of conceptions of patient's survival, cancer-related surgery, radiotherapy, and chemotherapy appeared to gradually increase from 2013 to 2015. This observation may be explained by the need for a longer follow-up period for studies on survival and therefore, only until recently, the longitudinal dataset of the NHIRD has accumulated a sufficient number of cancer cases for such evaluation. Rare cancers or risk factors with long induction periods will become feasible for investigation as the length of follow-up of the NHIRD cohort increases over time.

Our analyses of citation frequency of the articles indicated that the most cited study (92 times by PubMed Central articles as of July 30, 2016) was the one on the associations between nucleoside analog (e.g., Lamivudine, Adefovir, Entecavir) use and a lower risk of hepatocellular carcinoma recurrence among patients with hepatitis B virus-related hepatocellular carcinoma after liver resection.[9] A possible explanation for its high citation is that hepatocellular carcinoma is one of the leading causes of

Table 4

The 62 medical conceptions with at least 4 connections to other conceptions, generated from keywords identified in the 589 cancer articles based on the National Health Insurance Research Database between 2002 and 2015.

Conception	Keywords	n (%)
1. Diabetes	Diabetes, diabetogen, hyperglycemia, hyperosmolar, hypoglycemia, hypoglycaemia	118 (20.0)
2. Survival	Survival	113 (19.2)
3. Breast cancer	Breast cancer	96 (16.3)
4. Lung cancer	Lung cancer	67 (11.4)
5. colorectal cancer	Colorectal cancer	61 (10.4)
6. Cost	Cost, expenditure, financial, money	57 (9.7)
7. Liver cancer	Liver cancer, hepatoma, hepatocellular cancer, hepatocarcinoma, hepatoblastoma	49 (8.3)
8. Surgery	Surgery, operation	44 (7.5)
9. Prostate cancer	Prostate cancer	44 (7.5)
10. Chemotherapy	Chemotherapy or chemotherapies	41 (7.0)
11. Hypertension	Hypertension	41 (7.0)
12. Cervical cancer	Cervical cancer	35 (5.9)
13. Hyperlipidemia	Hyperlipidaemia, hyperlipidemia, hypertriglyceridemia, dyslipidaemia, dyslipidemia, cholesterol, gemfibrozil, ezetimibe, simvastatin, atorvastatin, rosuvastatin, lovastatin, fluvastatin, pravastatin	32 (5.4)
14. Radiotherapy	Radiotherapy	30 (5.1)
15. Diabetes medications	Sulfonylurea, sulphonylurea, sulfonamide, α reductase, PPARα, α glucosidase, thiazolidinedion, gastrointestinal, saxagliptin, pioglitazone, sitagliptin, rosiglitazone, glitazone, acarbose, insulin, DPP-4	30 (5.1)
16. Lymphoma	Lymphoma	27 (4.9)
17. Gastric cancer	Gastric cancer	25 (4.2)
18. Head and neck cancer	Head and neck cancer	24 (4.1)
19. Digestive diseases	Jaundice, gallstone, pylori, ultrasound, reflux	23 (3.9)
20. End of life	End of life	22 (3.7)
21. Chemotherapy medications	Fluorouracil, tamoxifen, doxorubicin, gefitinib	22 (3.7)
22. Stroke	Stroke	22 (3.7)
23. Thyroid cancer	Thyroid cancer	22 (3.7)
24. Renal or urinary cancer	Renal or urinary cancer	21 (3.6)
25. Hospice	Hospice	19 (3.2)
26. Cirrhosis	Cirrhosis, cirrhotic	19 (3.2)
27. Gynecology	Hysterectomy, eclampsia, estrogen, menopause, endometriosis, placenta, hysterosalpingography, infertility	18 (3.1)
28. Leukemia	Leukemia	18 (3.1)
29. Esophageal cancer	Esophageal cancer	18 (3.1)
30. Infarction	Stent, macrovascular, thrombolysis, ischemia, ischaemia, PAOD, infarction	17 (2.9)
31. Ovarian cancer	Ovarian cancer	15 (2.5)
32. Tuberculosis	Tuberculosis, antituberculosis, anti-TB	14 (2.4)
33. Asthma	Asthma, COPD, bronchiectasis	14 (2.4)
34. PAP	Pap	14 (2.4)
35. HBV or HCV	HBV, HCV	13 (2.2)
36. Cardiology	Cerebrovascular, vascular, myocardial cardiopulmonary, thromboembolic, dysrhythmia, arrhythmia, tachycardia, fibrillation, angina, renin, atherosclerotic, ventricular, aneurysm, troponin, aortic, cardiac, IABP	12 (2.0)
37. Cardiovascular medications	Amlodipine, atropine, bisoprolol, α blocker, alpha blocker, carvedilol, vasopressor, digoxin, β blocker, nicorandil, angiotensin converting enzyme, angiotensin receptor, angiotensin, candesartan, losartan	12 (2.0)
38. Anticoagulants	Antiplatelet, antithrombotic, anticoagulant, clopidogrel, clopidorel, dihydropyridine, cilostazol, aspirin, coumarin, warfarin, heparin	12 (2.0)
39. Obesity	Obesity	11 (1.9)
40. Endometrial cancer	Endometrial cancer	11 (1.9)
41. Surgeon volume	Surgeon volume	10 (1.7)
42. Chinese medicine	Acupuncture, tang, wan, saan, xiaofengsan, xinyiqingfeitang, xuefuzhuyutang, zhenrenhuomingyin, zhibaidihuangwan, guizhifulingwan, shangjongshiahtongyongtongfengwan, shenlingbaizhusan, shentongzhuyutang, haipiaoxiao, shinyiqingfeitang, shujinhuoxuetang, jiaweixiaoyaosan, liuweidihuangwang, maxingganshitang, jiaweixiaoyaosang, suanzaorentang, jishengshenqiwan, shujinghuoxietang, shujinghuoxuetang, danzhixiaoyaosan, guizhishaoyaozhimutang, yanhusuo, dangguiniantongtang, duhuojishengtang, buyanghuanwutang, buzhongyiqitang, xiangshaliujunzitang, xiaoqinglongtang, zhigancaotang, banxiaxiexintang, chuanlianzi, danzhixiaoyaosan, guipitang	10 (1.7)
43. Infection	Bacteremia, sepsis, septicemia, empyema, gonorrhea, enterococcus, salmonellosis, klebsiella, mycoplasma, cytomegalovirus, streptococcus, streptococci, scabi, scabies, antibiotics, antifungus, parasite	9 (1.5)
44. Palliative	Palliative	8 (1.4)
45. Pancreatic cancer	Pancreas cancer	8 (1.4)
46. Neurology	Parkinson, epilepsy, Alzheimer's, Alzheimer, hemiparesis, hemiplegia, vertigo, dementia, Huntington, subarachnoid, ataxia, seizure, headache, palsy	7 (1.2)
47. Blood	Hemoglobin, haemophilia, hemophilia, thalassaemia, thalassemia, anaemia, anemia	7 (1.2)
48. Toxic	Poison, arsenic, cholinesterase, anticholinesterase, hypoxia, heroin, homicide, suicidal, suicide	7 (1.2)
49. Dental	Periodontal, periodontitis	6 (1.0)

(continued)

Table 4
(continued).

Conception	Keywords	n (%)
50. Trauma	Lumbargo, fracture, traumatic, trauma, traffic, pneumothorax, hemopneumothorax	6 (1.0)
51. Colitis	Crohn disease, ulcerative colitis, inflammatory bowel disease	5 (0.8)
52. Pneumonia	Pneumonia, empyema, bronchitis, bronchopneumonia, pneumococcal	5 (0.8)
53. Dialysis or uremia	Dialysis, uremia	5 (0.8)
54. Thyroid diseases	Thyroiditis, Graves disease, Hashimoto	4 (0.7)
55. Lupus	Systemic lupus erythematosus, immune thrombocytopenia	4 (0.7)
56. Arthritis	Rheumatoid arthritis, juvenile idiopathic arthritis	4 (0.7)
57. Muscle-related autoimmune diseases	Guillain Barre syndrome, multiple sclerosis, myasthenia gravis	4 (0.7)
58. Gout	Benzbromarone, probenacid, allopurinol, anti hyperuricemia, antigout, hyperuricemia	4 (0.7)
59. Heart rate	Dysrhythmia, arrhythmia, tachycardia, fibrillation	4 (0.7)
60. Labor	Labor	4 (0.7)
61. Myeloma	Myeloma	4 (0.7)
62. Adenosquamous cancer	Adenosquamous cancer	4 (0.7)

COPD = chronic obstructive pulmonary disease, DPP-4 = dipeptidyl peptidase-4, HBV = hepatitis B virus, HCV = hepatitis C virus, IABP = intra-aortic balloon pump, PAOD = peripheral arterial occlusive disease, pap = Papanicolaou test, PPARγ = peroxisome proliferator-activated receptor-γ.

death for cancer patients and therefore, its treatment is widely studied, which leads to its citation by other research articles. It should be noted that the citation frequency obtained in this study reflected only the number of PubMed Central articles, which are all full-text articles freely accessible to the public (https://www.ncbi.nlm.nih.gov/pmc/) rather than by all articles indexed by PubMed or by other databases such as Web of Science, Scopus, and Google Scholar.[20]

Despite the NHIRD is a nationwide, population-based dataset, it has a number of inherent limitations that hinder its use. While the NHIRD represents a cohort of approximately 1 million patients, there still may not be enough cases for the study of certain rare cancers and their associations with other diseases with a low prevalence or incidence. In addition, no information on cancer stage is available from the dataset. Important potential confounding variables, such as body mass

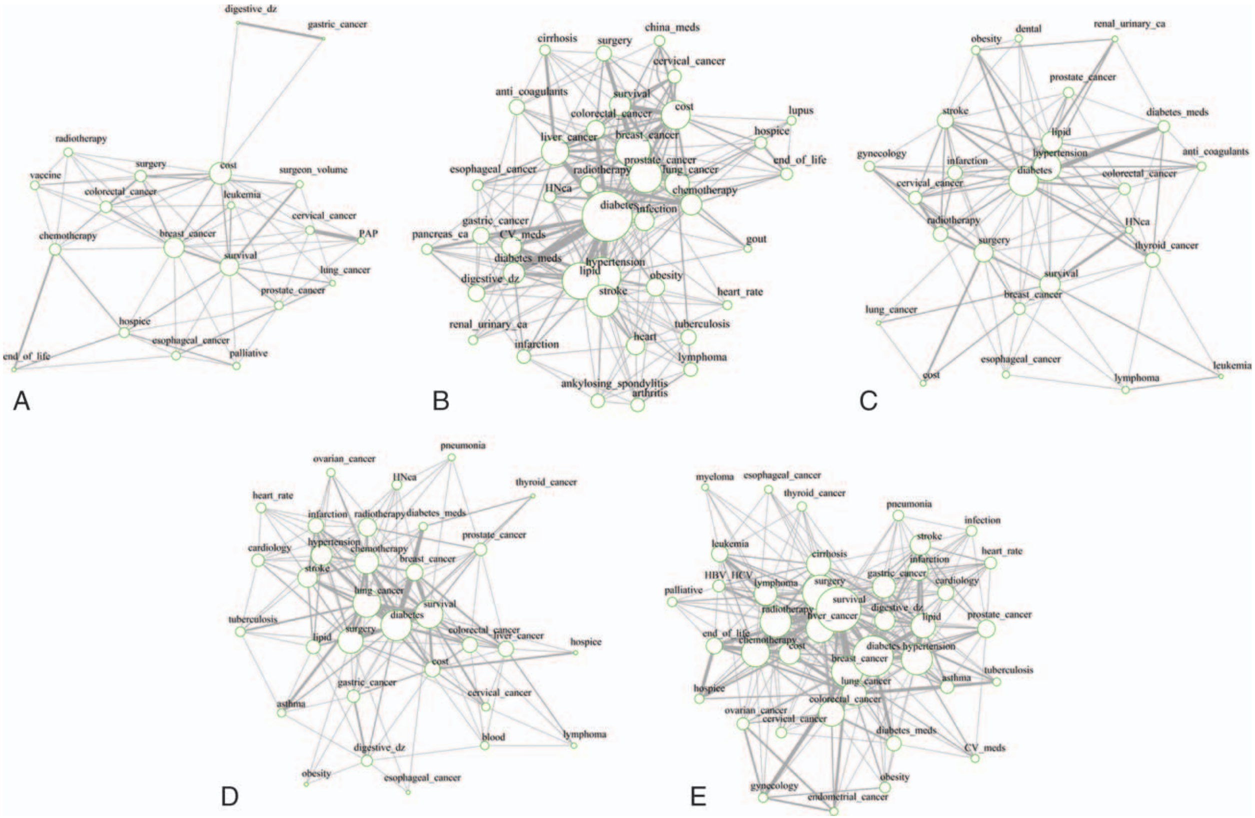

Figure 3. (A) Network plots of key conceptions in the abstracts of articles published between 2002 and 2010. (B) Network plots of key conceptions in the abstracts of articles published in 2011 and 2012. (C) Network plots of key conceptions generated from the abstracts of articles published in 2013. (D) Network plots of key conceptions generated from the abstracts of articles published in 2014. (E) Network plots of key conceptions generated from the abstracts of articles published in 2015. china_meds = Chinese medicine, CV_meds = cardiovascular medications, diabetes_meds = diabetes medications, digestive_dz = digestive diseases, HNca = head and neck cancer, pancreas_ca = pancreatic cancer, renal_urinary_ca = renal or urinary cancer.

index, smoking, and alcohol intake, are also not available from the dataset for any statistical adjustment. Moreover, the study of cancer medications is impeded by the lack of information on dosage and medication adherence. Furthermore, patients' use of self-paid medications or procedures is also not recorded in the NHIRD.

A few limitations of the present study should be mentioned. First, only articles written in English and indexed by PubMed were included in the study. Nevertheless, most medical research articles based on NHIRD should have been identified since researchers' performance in Taiwan is generally evaluated based on output of articles published in the Science Citation Index,[21] which is most likely to be covered by Medline. Second, articles with no abstract could not be analyzed. Third, variations in the length of abstract and inaccuracy in the content of the abstract might potentially influence our results.[22]

In conclusion, in this study of 589 published articles on secondary data analysis of the NHIRD, indexed by PubMed between 2002 and 2015, we found that the top 5 most studied cancers were breast, lung, colorectal, liver, and prostate. Articles generally focused on the association between cancer and its possible risk factors, such as diabetes, hypertension, dyslipidemia, and statin use. The conceptions of patients' survival, cancer-related surgery, radiotherapy, and chemotherapy were gradually increased from 2013 to 2015. Overall, there were more articles focusing on the risk factors of cancer, treatment of cancer, and survival of cancer patients, but relatively few articles on end-of-life cancer care including hospice care, palliative care, and home palliative care. These latter neglected areas should further be explored using the NHIRD as they are as important as treatment of the disease itself for many patients.

References

[1] Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359–86.

[2] Global Health Observatory data repository. Number of deaths (World) by cause. Estimates for 2000–2012. Available from: http://apps.who.int/gho/data/node.main.CODWORLD?lang=en. Accessed February 8, 2017.

[3] Cheng HG, Phillips MR. Secondary analysis of existing data: opportunities and implementation. Shanghai Arch Psychiatry 2014;26:371–5.

[4] National Health Insurance Administration, Ministry of Health and Welfare. 2015–2016 National Health Insurance Annual Report; 2015. Available from: http://www.nhi.gov.tw/Resource/webdata/13767_1_2015-2016%20NHI%20ANNUAL%20REPORT.pdf. Accessed February 8, 2017.

[5] National Health Research Institutes. National Health Insurance Research Database, data subsets. Available from: http://nhird.nhri.org.tw/en/Data_Subsets.html. Accessed November 9, 2016.

[6] Chen YC, Yeh HY, Wu JC, et al. Taiwan's National Health Insurance Research Database: administrative health care database as study object in bibliometrics. Scientometrics 2011;86:365–80.

[7] U.S. National Library of Medicine. MEDLINE, PubMed, and PMC (PubMed Central): How are they different? 2016. Available from: https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html. Accessed February 8, 2017.

[8] Castillo M. Is your journal indexed in MEDLINE? Am J Neuroradiol 2011;32:1–2.

[9] Wu CY, Chen YJ, Ho HJ, et al. Association between nucleoside analogues and risk of hepatitis B virus-related hepatocellular carcinoma recurrence following liver resection. JAMA 2012;308:1906–14.

[10] Wu CY, Kuo KN, Wu MS, et al. Early *Helicobacter pylori* eradication decreases risk of gastric cancer in patients with peptic ulcer disease. Gastroenterology 2009;137:1641–8.

[11] Lai MN, Wang SM, Chen PC, et al. Population-based case-control study of Chinese herbal products containing aristolochic acid and urinary tract cancer risk. J Natl Cancer Inst 2010;102:179–86.

[12] Tsan YT, Lee CH, Ho WC, et al. Statins and the risk of hepatocellular carcinoma in patients with hepatitis C virus infection. J Clin Oncol 2013;31:1514–21.

[13] Chiu HF, Ho SC, Chen CC, et al. Statin use and the risk of liver cancer: a population-based case-control study. Am J Gastroenterol 2011;106: 894–8.

[14] Lin CC, Chiang JH, Li CI, et al. Cancer risks among patients with type 2 diabetes: a 10-year follow-up study of a nationwide population-based cohort in Taiwan. BMC Cancer 2014;14:381.

[15] Lai SW, Liao KF, Chen PC, et al. Antidiabetes drugs correlate with decreased risk of lung cancer: a population-based observation in Taiwan. Clin Lung Cancer 2012;13:143–8.

[16] Tseng CH. Pioglitazone and bladder cancer: a population-based study of Taiwanese. Diabetes Care 2012;35:278–80.

[17] Chang CH, Lin JW, Wu LC, et al. Association of thiazolidinediones with liver cancer and colorectal cancer in type 2 diabetes mellitus. Hepatology 2012;55:1462–72.

[18] Chang CH, Toh S, Lin JW, et al. Cancer risk associated with insulin glargine among adult type 2 diabetes patients—a nationwide cohort study. PLoS ONE 2011;6:e21368.

[19] Taiwan Cancer Registry. Cancer Incidence and Mortality Rates in Taiwan. Available from: http://tcr.cph.ntu.edu.tw/main.php?Page=N2. Accessed February 14, 2017.

[20] Kulkarni AV, Aziz B, Shams I, et al. Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. JAMA 2009;302:1092–6.

[21] Wu YT, Lee HY. National Health Insurance database in Taiwan: a resource or obstacle for health research? Eur J Intern Med 2016;31: e9–10.

[22] Fontelo P, Gavino A, Sarmiento RF. Comparing data accuracy between structured abstracts and full-text journal articles: implications in their use for informing clinical decisions. Evid Based Med 2013; 18:207–11.