Journal of Informetrics 10 (2016) 273–285

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Trajectory analysis of drug-research trends in pancreatic cancer on PubMed and ClinicalTrials.gov

Yoo Kyung Jeong[a], Go Eun Heo[a], Keun Young Kang[a], Dong Sup Yoon[b], Min Song[a,*]

[a] Department of Library and Information Science, Yonsei University, Republic of Korea
[b] Pancreatobiliary Cancer Clinic, Department of Surgery, Gangnam Severance Hospital, Yonsei University College of Medicine, Republic of Korea

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form 9 January 2016
Accepted 10 January 2016
Available online 11 February 2016

Keywords:
Pancreatic cancer
Text mining
Bibliometric analysis
Data analysis
Information extraction

ABSTRACT

Increasing interest in developing treatments for pancreatic cancer has led to a surge in publications in the field. Analyses of drug-research trends are needed to minimize risk in anti-cancer drug development. Here, we analyzed publications on anti-cancer drugs extracted from PubMed records and ClinicalTrials datasets. We conducted a drug cluster analysis by proposing the entity Dirichlet Multinomial Regression (eDMR) technique and in-depth network analysis of drug cluster and target proteins. The results show two distinct research clusters in both the ClinicalTrials dataset and the PubMed records. Specifically, various targets associated with anti-cancer drugs are investigated in new drug testing while the diverse chemicals are studied together with a standard therapeutic agent in the academic literature. In addition, our study confirms that drug research published in PubMed is preceded by clinical trials. Although we only evaluate drugs for pancreatic cancer in the present study, our method can be applied to drug-research trends of other diseases.

1. Introduction

Pancreatic cancer is hard to diagnose and is often accompanied by a poor prognosis with a very low survival rate. According to the National Cancer Institute (NCI), only 6.7% of people survive 5 years or more after being diagnosed with pancreatic cancer (NCI, 2012). Chemotherapy is the most common treatment. Due to the limited number of drugs, doctors often prescribe multiple chemotherapy agents, and much recent research has focused on finding an optimal drug combination (NCI, 2012).

Increasing investments in pancreatic cancer research have driven a surge in relevant publications. As of February 20, 2015, the query "pancreatic cancer" returned 73,771 records from the PubMed database. To understand the field, researchers adopted bibliometric approaches to identify the knowledge structure of a target discipline by analyzing bibliographic metadata such as authors, institutes, and countries (Lewison, Purushotham, Mason, McVie, & Sullivan, 2010; López-Illescas, de Moya-Anegón, & Moed, 2008; Mela, Cimmino, & Ugolini, 1999; Ugolini, Casilli, & Mela, 2002; Ugolini & Mela, 2003). Bibliometrics studies focused on oncology aim to understand changes and characteristics in the field by analyzing the productivity of different journals, papers, and authors (Lewison et al., 2010; López-Illescas et al., 2008; Mela et al., 1999; Ugolini et al.,

* Correspondence to: Min Song, Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea. Tel.: +82 2 2123 2416; fax: +82 2 393 8348.
E-mail address: min.song@yonsei.ac.kr (M. Song).

http://dx.doi.org/10.1016/j.joi.2016.01.003
1751-1577/

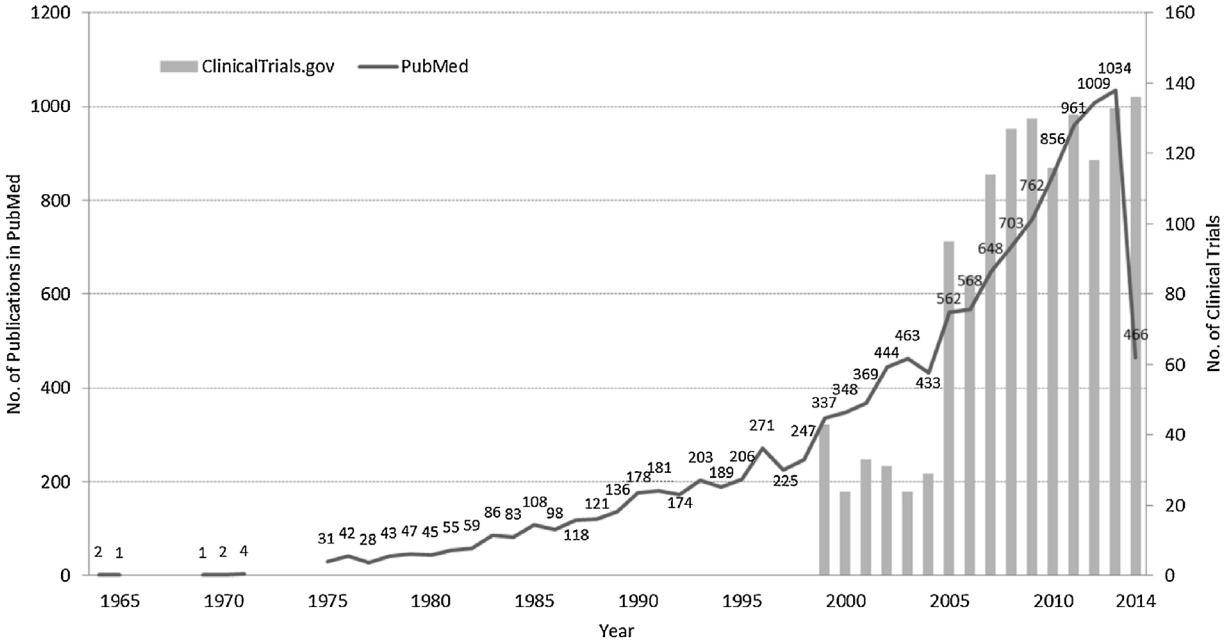

Fig. 1. Publication growth rate related to pancreatic cancer drugs.

2002; Ugolini & Mela, 2003). Mela et al. (1999) and Ugolini et al. (2002) identified popular research topics by conducting keywords analysis of oncology journals published during 1995 in the European Union (EU). They also compared the results of the previous study by considering papers published between 1996 and 2000 (Ugolini & Mela, 2003). In a citation-impact analysis, López-Illescas et al. (2008) analyzed oncological research articles and compared bibliometric indicators such as article counts, journal impact factors, and actual citation counts. More recently, Lewison et al. (2010) evaluated cancer research at United Kingdom Cancer Centres (UKCC) by measuring the citation impact. However, these studies only analyzed overall cancer research fields in a quantitative manner rather than analyzed the topical trends or clinical implications. Recently, researchers have adopted text-mining techniques to make sense of many publications through the lens of content analysis (Jeon et al., 2014; Zhao & Weng, 2011). Zhao and Weng (2011) proposed a pancreatic cancer prediction model by analyzing text in PubMed and EHR data. For discovering novel drug targets, Jeon et al. (2014) used text-mining techniques to identify cancer-associated proteins.

The goal of the present study is to understand the complete portrait of anti-cancer drug research by analyzing the anti-cancer drug clusters extracted both from scholarly publications and clinical trials. In particular, we investigate whether knowledge transfer occurs between scholarly publications and clinical trials. For clinical trials, we use ClinicalTrials.gov, a repository for shared clinical trials results that provides the ability to track journal articles about each trial. For scholarly publications, we use PubMed, a database of more than 22 million published articles. We discover drug clusters using topic-modeling technique and trace changes of drug targets associated with anti-cancer drugs with the aid of the target-drug network. In addition, we investigated topical resemblance or discrepancy of the anti-cancer drugs investigated in scientific publications versus clinical trials. Comparison of these two distinct datasets gives us a new insight into anti-cancer drug research in laboratory and clinical settings. For trajectory drug cluster analysis, we propose the entity Dirichlet-Multinomial Regression (eDMR) topic-modeling technique, which is an extension of DMR (Mimno & McCallum, 2012) to identify the topical trends of anti-cancer drug research over time. We also analyze the target proteins of anti-cancer drugs in the drug network to provide a macro level of view for tracing research trends in pancreatic cancer.

The present paper is organized as follows. In Section 2, we describe the datasets obtained from two different sources and the experiment design and procedure. In Section 3, we explain and analyze the results of topic-model and network analysis. In the last section, we summarize the findings and discuss implications and suggestions for future work.

2. Methods

2.1. Data collection

Our primary goal is to identify the comprehensive landscape encompassing both scientific publications and clinical trials in pancreatic cancer drug. To this end, we collected 14,695 scholarly publications from PubMed and 3152 trials from ClinicalTrials.gov by searching these two data collections with the query term “pancreatic cancer drug.”

Table 1
Number of records in Clinicaltrials.gov and PubMed.

Rank	ClinicalTrials.gov	Frequency	PubMed	Frequency
1	Gemcitabine	497	Gemcitabine	3887
2	Capecitabine	110	Insulin	2200
3	Fluorouracil	100	Fluorouracil	1489
4	Oxaliplatin	89	Glucose	1045
5	Erlotinib	85	Somatostatin	1042
6	Placebo	67	Epidermal Growth Factor	925
7	Nab Paclitaxel	50	Octreotide	513
8	Cisplatin	47	Vascular endothelial growth factor A	512
9	Irinotecan	47	Hormones	475
10	Docetaxel	37	Cisplatin	454

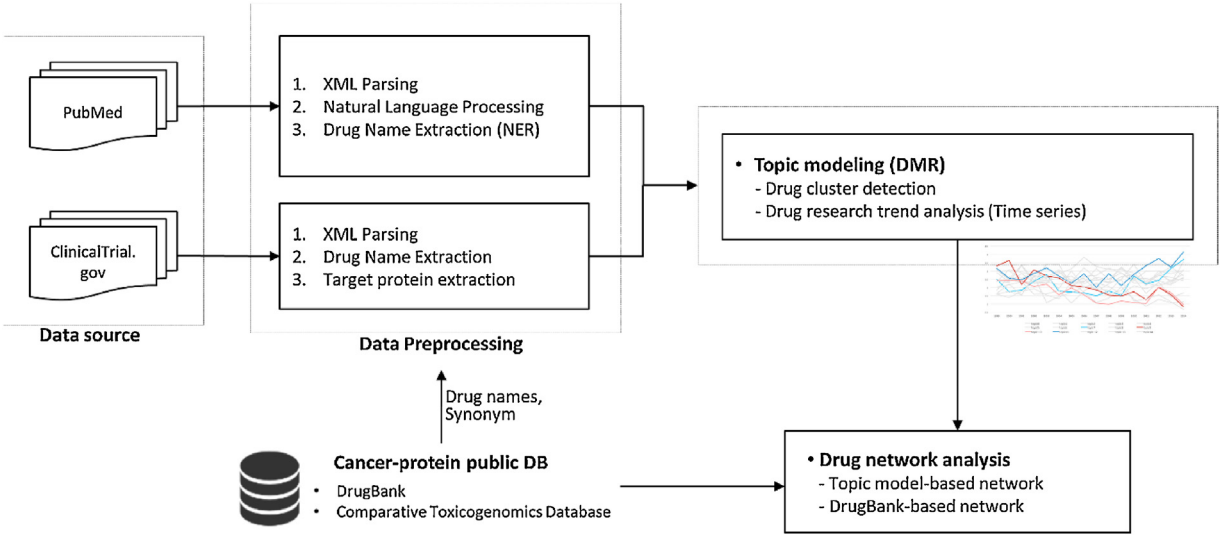

Fig. 2. Flowchart of research overview.

Fig. 1 shows the publication growth rate for pancreatic cancer drugs in PubMed and ClinicalTrials.gov. The number of scientific papers concerning pancreatic cancer increased in recent years with a similar pace to that of the clinical trials. However, after 2005, the number of PubMed-listed papers continuously increased, whereas the number of clinical trials increased about three times rapidly from 29 in 2004 to 95 in 2005. The FDA approval of the gemcitabine/erlotinib combination as a first-line therapy for pancreatic cancer in 2005 (PharmaCyte Biotechm, 2015) may be attributed to the growth in chemotherapy combination trials.

Unlike existing bibliometric studies in cancer, we performed entity extraction to detect drug names in PubMed papers and clinical trial records. We extracted 3618 unique drug names from PubMed and 810 from ClinicalTrials.gov. Table 1 shows top 10 drugs by frequency in PubMed and ClinicalTrials.gov data. Gemcitabine, a standard therapy of pancreatic cancer, is the most frequently mentioned drug in both databases. In PubMed, anticancer drugs and hormones are also highly ranked. Other drugs such as Nab Paclitaxel, Irinotecan, and Docetaxel are among the most common on ClinicalTrials.gov. The drugs such as gemcitabine, fluorouracil, capecitabine (the oral prodrug of fluorouracil) occur frequently in both clinical and PubMed data. These drugs, which hinder DNA synthesis in cancer cells, are used frequently as a first-line treatment for pancreatic cancer (Law et al., 2014). In PubMed, however, hormones like insulin are disproportionately common, compared with ClinicalTrials.gov. The pancreas secretes insulin when blood glucose levels rise. Secretion of the hormone can cause the cancer to grow (Gullo, Pezzilli, & Morselli-Labate, 1994).

2.2. Research overview

Fig. 2 shows our approach to analyzing drug-research trends. We collected data from PubMed and ClinicalTrials.gov in XML format, and we developed XML parsers to extract elements such as title and abstract from the listed papers and trials. We used the Stanford Core NLP (Manning et al., 2014) to preprocess texts such as titles and abstracts. Then, we employed the Named Entity Recognition (NER) technique provided by PKDE4J, a dictionary-based biomedical text mining tool (Song, Kim, Lee, Heo, & Kang, 2015), to extract drug names. PKDE4J is a flexible pipeline-based bio text mining tool for entity extraction and their relations from unstructured biomedical text. PKDE4J was used to extract entity names and avoid the problem of a high number of false positives. It was reported that PKDE4J has fairly good performance in accuracy with average F-measure

of 85% on several corpus such as GENETAG (Tanabe, Xie, Thom, Matten, & Wilbur, 2005), CRAFT (Bada et al., 2012), and NCBI Disease Corpus (Doğan, Leaman, & Lu, 2014). We used two dictionaries, the Comparative Toxicogenomics Database (CTD) (Davis et al., 2012) and DrugBank (Law et al., 2014), to cover as many drug (or chemical) names as possible.

2.3. eDMR: a drug clustering technique

After extracting drug names from two datasets, we conducted time-series analyses using an extension of the DMR topic-modeling technique (Mimno & McCallum, 2012), a variation of Latent Dirichlet allocation (LDA). Topic modeling is a generative statistical algorithm to discover topics (or word clusters) that occur in a collection of documents based on co-occurring words. Based on the probability of word occurrence and distribution in a collection of documents, topic models identify words that are found together at a disproportionately high rate. DMR incorporates the probability of a word over documents. DMR allows for conditioning on arbitrary document features by including a long-linear prior on document-topic distributions that is a function of the features of the document such as author, years, and references. We used a modified technique called eDMR specific for drug clustering. For each document d, let x_d be a feature vector representing drug and date. Given the prior distribution of $N(0, \Sigma)$ and a hyper-parameter β, the generative process for documents and drugs mentioned in the documents is as follows:

(1) For each drug cluster c, draw $\emptyset_t \sim \mathrm{Dir}(\beta)$ noting that $\mathrm{Dir}(\beta)$ is a distinct Dirichlet distribution with the Dirichlet prior on the topic-drug distribution, β.
(2) For each document d,

Draw $\theta_d \sim \mathrm{Dir}(\alpha_d) = \mathrm{Dir}(\exp(\tau_d))$ with $\tau_d \in \tau$ noting a per-document α_d, the parameters of a Dirichlet distribution and τ_d are a covariance function $f(a_d, b_k)$, where a_d is the observed attribute vector of document d and b_k is a vector of drug and date.

For each drug du,

Draw $Z_{d,du} \sim \mathrm{Multi}(\theta_d)$ to denote that $Z_{d,du}$ is the cluster assignment of a drug t_{du} and θ_d is cluster proportion of a document d.

Draw $T_{d,du} \sim \mathrm{Multi}(\emptyset_{Z_{d,du}})$ to denote that $T_{d,du}$ is the du-th drug of a document d and $\emptyset_t$ is preference of a cluster t over the vocabulary with $\sum_n \emptyset_{t,n} = 1$.

For eDMR drug clustering, we set three fixed parameters: σ^2, the variance of the prior on parameter values for prior distribution; β, the Dirichlet prior on the cluster-drug distributions; and $|T|$, the number of clusters.

Through this process, we can obtain drug (or chemical) clusters from each dataset and identify the trends in drug research for pancreatic cancer. For in-depth analysis, we constructed two networks of drug clusters. The networks represent the relationship between drugs and its clusters in two perspectives. One network is created with eDMR results by using co-occurrence frequencies as edge weight; the other is constructed from DrugBank (Law et al., 2014) by using a target as an edge of two drugs linked by the target protein. For visualization of drug network, we use Gephi, an open source visualization tool.

3. Results

3.1. Drug cluster analysis

Because we used year as the third parameter in our evaluation of drug clusters, eDMR enabled time-variant analysis of research trends about pancreatic cancer drugs. To test the approach, we varied the number of topics (10, 15, 20, or 25) and analyzed each result manually. We observed that 15 topics are best representing the drug clusters.

eDMR results are presented in two formats: drug clusters (See Tables 2 and 3) and chronological graphs (See Figs. 3 and 4). Each drug cluster is labeled and drug names within each topic are sorted by statistical probability of a drug in the drug cluster that denotes the representativeness of the drug for the given cluster.

Table 2 shows labeled drug clusters and the top five drugs within each cluster. A surgeon specializing in pancreatic cancer labels clusters and selects 12 most important ones among 15 drug clusters. The clusters in Table 2 can be grouped into four subtypes: Agent, Immuno-suppressants, Inhibitors, and Therapeutics. Agent clusters (Clusters 0, 1, and 2) are pertinent to chemotherapy drugs. Cluster 0 consists of standard therapeutic agents, such as gemcitabine, capecitabine, cisplatin, and fluorouracil, which can be given individually to patients. Standard agents can also combined with target agents, such as erlotinib or capecitabine. Some new therapies, such as folifirinox, are combination therapies. Folifirinox consists of four anticancer drugs: leucovorin, fluorouracil, irinotecan, and oxalipaltin (Conroy, Gavoille, Samalin, Ychou, & Ducreux, 2013).

Fig. 3 depicts drug trends over the last 15 years. Among 15 clusters, four clusters (6, 8, 11, and 13) show noticeable changes. Clusters 8 (stomach-cancer therapeutics) and 11 (immunosuppressants) show the decreasing pattern in clinical trial fields (red lines). Before 2000, Cluster 8 rose steeply, but after 2000 the cluster declined. Cluster 11 has been unsteadily falling. Clusters 8 and 11 show a wide gap in the trend between 2002 and 2008. Clusters related to new drugs showed a distinct drug-research trend in clinical trials. Clusters 6 (new-drug testing) and 13 (modified anticancer drugs) show the steadily

Table 2
eDMR-based drug clustering results—ClinicalTrials.gov.

	Drug 1	Drug 2	Drug 3	Drug 4	Drug 5
Cluster 0 (Standard therapeutic agent)	Gemcitabine	Capecitabine	Cisplatin	Docetaxel	Fluorouracil
Cluster 1 (New agent)	FOLFIRI-NOX	LDE225	MEDI-565	NVP-BKM120	Ascorbic acid
Cluster 2 (Target agent)	Gemcitabine	Erlotinib	Oxaliplatin	Capecitabine	Sorafenib
Cluster 3 (Immuno-suppressant)	Cyclophospha-mide	M2Es	Synthetic human secretin	RO4929097[a]	Gamma-secretase
Cluster 4 (Adenocarcinoma)	Placebo	Ciprofloxacine	Capecitabine	INCB018424	d-Methadone
Cluster 5 (Inhibitors)	Everolimus	Octreotide	Hhantag691	Vatalanib	2-carboxamide[b]
Cluster 6 (New drug testing)	Gemcitabine	Metformin	RTA402	Placebo	AMG479
Cluster 7 (New drug)	Gemcitabine	Bosutinib	Paclitaxel albumin[c]	Hydroxychloro-quine	Temsirolimus
Cluster 8 (Stomach cancer therapeutics)	Gemcitabine	s-1	Paclitaxel	Carboplatin	GSK1120212
Cluster 9 (Colon cancer therapeutics)	Fluorouracil	Oxaliplatin	Irinotecan	Leucovorin	Leucovorin calcium
Cluster 10 (Breast cancer)	Tipifarnib	Cyclophospha-mide	Fluorouracil	Doxorubicin	Rubitecan
Cluster 11 (Immuno-suppressant)	Sirolimus	Tacrolimus	Mycrophenolate mofetil	Enoxaparin	Pf00562271
Cluster 12 (Neuroendocrine)	Sunitanib	Ketoconazole	90y-hpam4	LCT	MCT
Cluster 13 (Modified anti-cancer drugs)	Gemcitabine	130-nm albumin-bound paclitaxel	Albumin-bound paclitaxel	Anti-interleukin[d]	CNTO 328
Cluster 14 (DNA enzyme inhibitor)	Placebo	Olaparib	Levofloxacin	MP-376	Morab-009

[a] Notch Signalling Pathway Inhibitor RO4929097.
[b] 2-((r)-2-methylpyrrolidin-2-yl)-1h-benzimidazole-4-carboxamide.
[c] Paclitaxel albumin-stabilized.
[d] Anti-interleukin-6 monoclonal antibody.

Table 3
eDMR-based drug clustering results—PubMed.

	Drug 1	Drug 2	Drug 3	Drug 4	Drug 5
Cluster 0 (Beta cell related)	Insulin	Glucose	Glucagon	Calcium	Hormones
Cluster 1 (Cytokines)	Nitric Oxide	Cytokines	Interferons	Interleukin-1	Tamoxifen
Cluster 2 (Anti-tumor effect)	Mitomycin	Doxorubicin	Celecoxib	Fluorouracil	Cholesterol
Cluster 3 (Adjuvant agent)	Curcumin	Nicotine	Emodin	Cadherins	Zinc
Cluster 4 (Target agent)	Epidermal growth factor	Erlotinib	Bortezomib	Ifosfamide	Gefitinib
Cluster 5 (Chemo agent)	Gemcitabine	Fluorouracil	Oxaliplatin	Capecitabine	Cisplatin
Cluster 6 (Antioxidant)	Somatostatin	Octreotide	Hormones	Transforming growth factor β	VIP[a]
Cluster 7 (Anti-folate)	Tumor necrosis factor-alpha	Pemetrexed	Folic acid	Phorbol	NS-398[b]
Cluster 8 (Tumor growth inhibition effect)	Reactive oxygen species	Metformin	Oxygen	Superoxides	Glutathione
Cluster 9 (Hormones)	Gastrins	Cholecystokinin	Peptides	Tretinoin	Bombesin
Cluster 10 (Antiproliferative agent)	VEGF-A[c]	Everolimus	Sunitinib	Sirolimus	Sorafenib
Cluster 11 (Anti-cancer agent)	Streptozocin	Ethanol	Rubitecan	Dexamethasone	Triptolide
Cluster 12 (Antigens)	Antigens	Glucagon-like peptide 1	Sodium	Thymidine	Hepatocyte Growth Factor
Cluster 13 (Anti-cancer agent)	Fluorouracil	Cisplatin	Irinotecan	Leucovorin	Paclitaxel
Cluster 14 (Neurotransmitter)	Calcium	Melatonin	Adenosine triphosphate	Adenosine	Pancreastatin

[a] Vasoactive Intestinal Peptide.
[b] n-[2-(cyclohexyloxy)-4-nitrophenyl]methanesulfonamide.
[c] Vascular endothelial growth factor A.

rising pattern (blue lines). Overall, these results indicate that new drugs are actively developed and tested by modifying existing anti-cancer drugs. The results also show that immunosuppressants have become a less popular research subject in recent years.

Compared with clinical data, PubMed records indicate more interest in the role or the effect of drugs in patients than in clinical outcomes (see Table 3). Clusters 2 (antitumor effect) and 8 (tumor growth–inhibition effect) are related to drug effect. Mitomycin, doxorubicin, and fluorouracil are anticancer drugs. Cluster 8 includes antioxidants. Clusters 3, 5, 10, 11, and 13 are associated with agents. Cluster 3 is an adjuvant agent that helps boost immunity. Curcumin has anticancer potential to suppress proliferation of a wide variety of tumor cells (Aggarwal, Kumar, & Bharti, 2003), and emodin inhibits cell adhesion of various human cancer cells (Huang, Shen, Shui, Wenk, & Ong, 2006), which is related with progression of cancer (Paul, Ewing, Jarrard, & Isaacs, 1997). Cluster 5 is a chemotherapeutic agent used in combination with standard therapeutics agents.

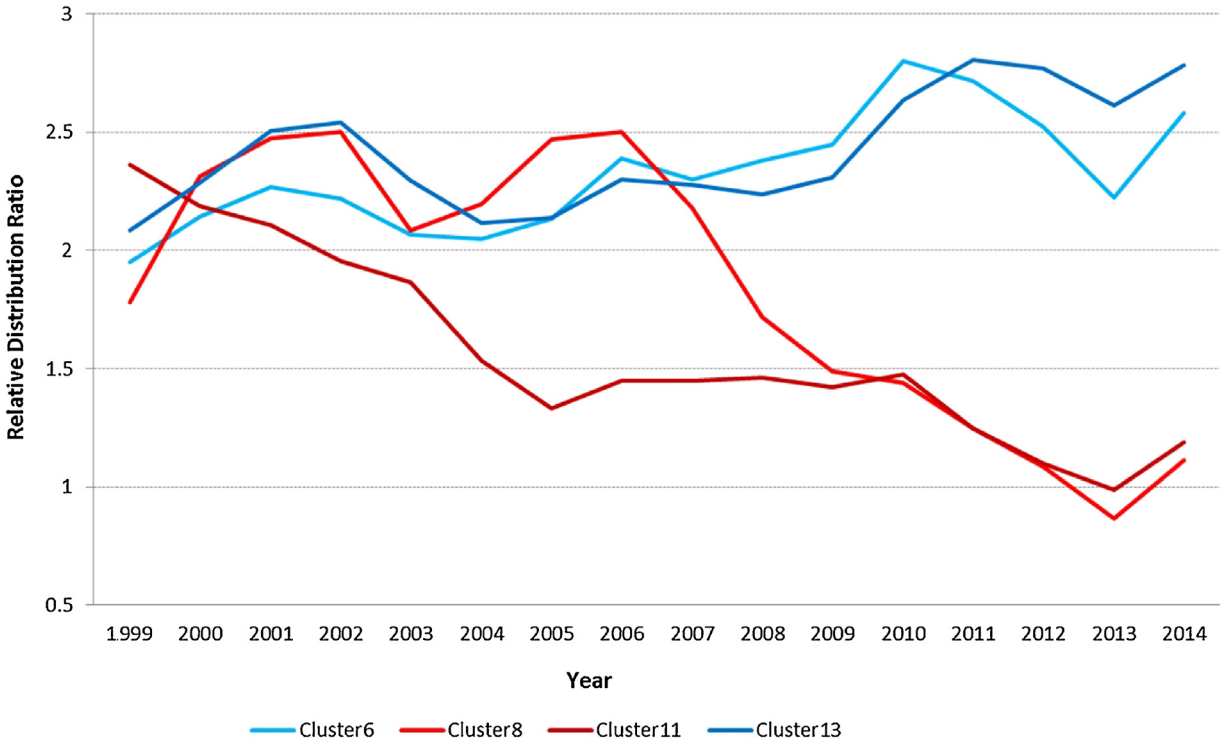

Fig. 3. Overall topic trends in clinical data.

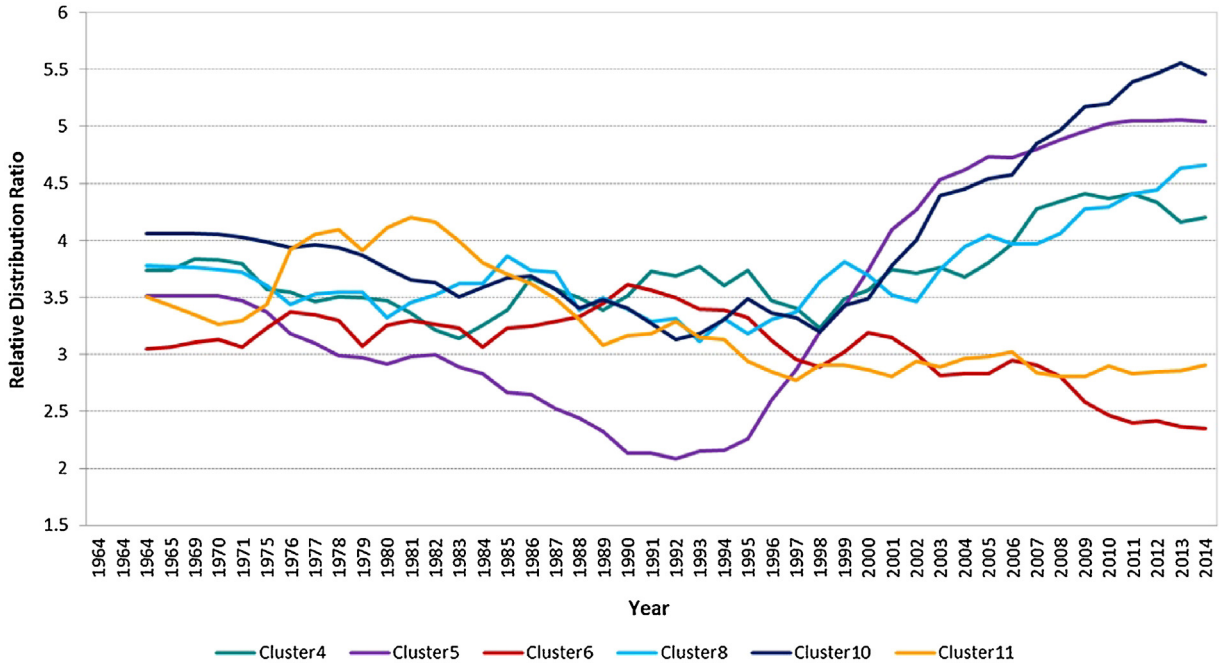

Fig. 4. Overall topic trends in PubMed data.

Cluster 10 is an antiproliferate agent and inhibits proliferation of cancer. Clusters 11 and 13 are related to anticancer agents and deal with components that are usually used in anticancer drugs.

Clusters related to agent (Clusters 4, 5, 8, and 10) show the rising trend (see Fig. 4), while research on anticancer drugs and antioxidants is decreasing. After 2000, agent-related clusters rose steadily. Target agents (Cluster 4) help intensive treatment of a specific cancer. Epidermal growth factor, in Cluster 4, is related to cell growth and is closely associated with the cancer's

Table 4
Cluster similarity within databases.

ClinicalTrial.gov			PubMed		
Cluster	Cluster	Similarity	Cluster	Cluster	Similarity
C 6	C 7	0.917	C 7	C 1	0.169
C 2	C 7	0.815	C 2	C13	0.159
C 7	C13	0.783	C 0	C14	0.135
C 2	C 6	0.775	C 5	C13	0.067
C 0	C 7	0.752	C 6	C 9	0.055
C 6	C13	0.744	C 7	C14	0.046
C 4	C14	0.720	C 1	C12	0.043
C 0	C 6	0.716	C 6	C10	0.025
C 7	C 8	0.715	C 3	C12	0.020
C 0	C 2	0.711	C 4	C10	0.019
Average similarity		0.123	Average similarity		0.010

Table 5
Cluster similarity between databases.

PubMed\ClinicalTrial.gov	C0	C2	C5	C6	C7	C8	C9	C10	C12	C13
C2	–	–	–	–	–	–	–	0.221	–	–
C5	0.799	0.847	–	0.929	0.977	0.731	–	–	–	0.801
C10	–	–	0.248	–	–	–	–	–	0.254	–
C11	–	–	–	–	–	–	–	0.303	–	–
C13	0.221	–	–	–	–	–	0.733	0.367	–	–

tissue cell division. Cluster 5 includes standard drugs and chemotherapy agents and has overlap with the drugs of Cluster 0 (standard therapeutics agent). Clusters 4 and 10 are related to antiproliferative agents, which block the over-proliferation of cancer cells and suppress tumor angiogenesis. Among drugs in Cluster 10, vascular endothelial growth factor A is a vital factor in angiogenesis. The other drugs, such as everolimus and sorafenib, inhibit of angiogenesis. Cluster 8 relates to the inhibition of tumor growth and oxygen-related medicine. Oxygen is a basic element in angiogenesis and cell growth. To inhibit the tumor growth, suppressing reactive oxygen species is needed.

Clusters 6 (antioxidant) and 11 (anticancer agent) show a downward trend. Cluster 6, which relates to antioxidants and drugs that target cancer-cell angiogenesis and proliferation, had low popularity in the late 1990s and early 2000s, but is less common now. Drugs in this cluster include streptozocin, ethanol, rubitecan, dexamethasone, and triptolide, agents that have not been commonly used since the 1980s.

To compare the difference in drug clusters obtained from each data set, we calculated the similarity of top 50 drugs for all 15 clusters. We use cosine similarity for measuring the likeness of clusters, and drugs are weighed by Term Frequency*Inverse Document Frequency (TF*IDF). We computed cluster similarity both within- and between databases.

Table 4 shows top 10 cluster pairs and their similarities within each data set. In clinical data, the degree of similarity is generally high. The average of similarity in clinical trials is 0.123, larger than the average similarity of PubMed 0.010. It might be caused by the limitation of drug usage in chemotherapy combination trials, while PubMed included various drug names even not tested in their articles. In fact, the largest similarity value is 0.917 (Clusters 7 and 6) in dataset from ClinicalTrial.gov. In particular, Cluster 7 (new drugs) has the highest similarity value with other clusters (0, 2, 6, and 13). It implies the trends of clinical testing, which combines new drugs with existing drugs in chemotherapy.

We also computed cluster similarity between clinical trials and PubMed records. Table 5 represents cluster pairs that exceed a similarity score of 0.2. Cluster 5 in PubMed (chemotherapy agents) has a high similarity value with the six clinical clusters (0, 2, 6, 7, 8, and 13). Since drugs used in chemotherapy have to be examined first by new drug testing, PubMed Cluster 5 has the highest similarity value (0.929) with the ClinicalTrials.gov Cluster 6 (new-drug testing). For the same reason, drugs in PubMed's Cluster 5 are generally studied in clinical trials.

To compare the datasets, we visualized drug clusters built from eDMR results. The assumption of constructing such a network is that drugs existing in the same cluster share similar characteristics. Each node in the network represents a drug; edges represent co-occurrence frequency between drugs. The size of nodes represents degree centrality (Wasserman & Faust, 1994), or the number of edges connecting that node. We use the modularity algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) for detecting drug clusters.

Fig. 5 shows the drug-cluster network of data from ClinicalTrials.gov. This network consists of 679 nodes and 16,624 edges. Among 15 clusters, drugs appearing in several clusters play a role in bridging the nodes. As shown in Fig. 6, Clusters 0, 2, 6, 7, and 13 deal with new drugs with high within-database similarity values.

Fig. 6 shows the drug cluster network from PubMed records, and the network has 529 and 15,826 edges. Each topic may belong to more than one community. Drugs in Cluster 12, for example, are also in Clusters 2 (antitumor effect), 5 (chemotherapy agent), and 13 (anticancer agent). Most drugs in Clusters 0 and 14 are found in both clusters; the two clusters have a within-database similarity score of 0.135. Clusters 1 and 7, which related to cytokines, also have a relatively

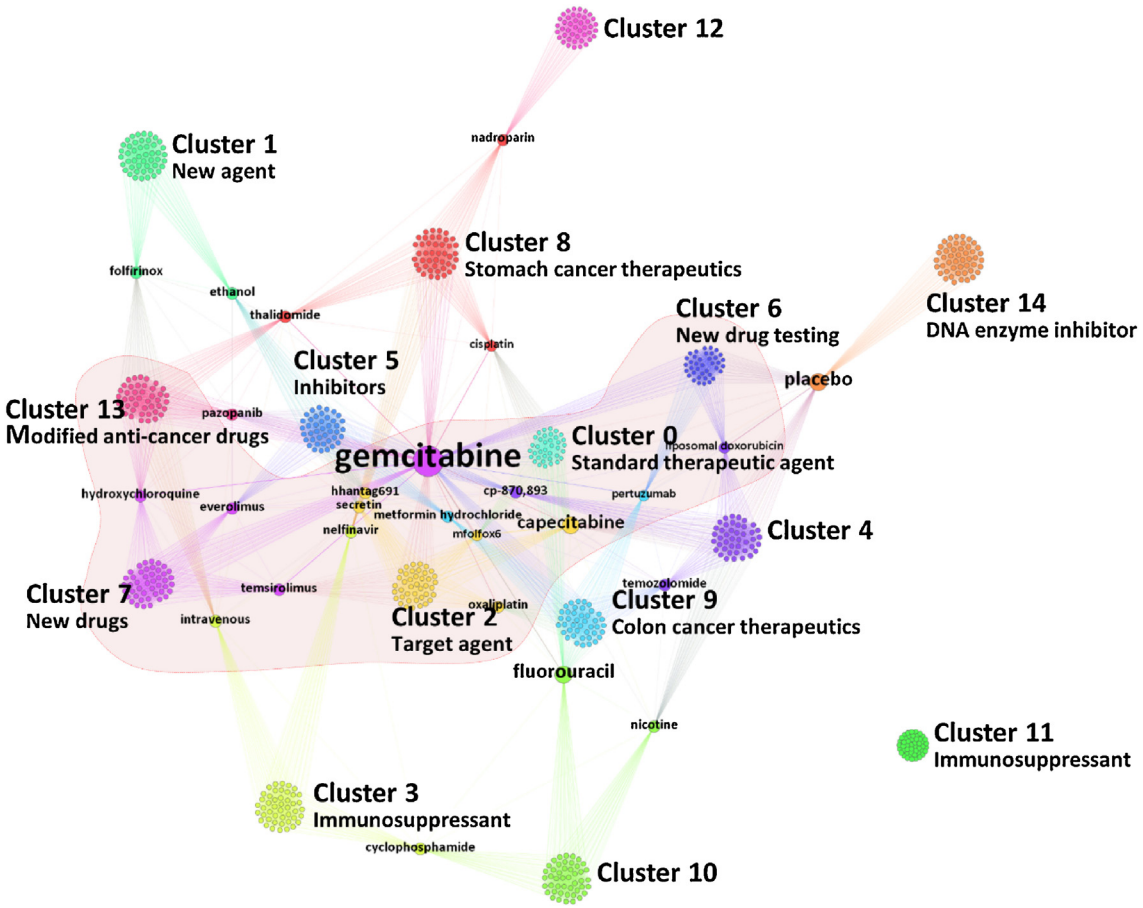

Fig. 5. ClinicalTrials.gov topic network.

high similarity value (0.169). The similarity scores are very low, however, in comparison to the scores of clusters in the ClinicalTrials.gov data set.

These results are consistent and confirm the drug clustering results that PubMed data covers a greater range of drugs than the data from ClinicalTrials.gov do. These drug-network analyses reveal that clinical trials and published articles provide two different perspectives for drug research. In clinical data, particular anticancer drugs are frequently referred to, while general chemicals such as sodium, phosphatidylinositols, and hormones are mentioned in PubMed data.

3.2. Target-based drug network analysis

For an in-depth analysis of drug clusters, we constructed the target-based drug network (TD network) using DrugBank (Law et al., 2014), which contains protein information (see Fig. 7). Using a similar approach, Yıldırım, Goh, Cusick, Barabási, and Vidal (2007) built a heterogeneous drug-target network and observed trends in drug design and differences between etiological and palliative drugs. In the present paper, we are primarily interested in how drug clusters overlap in the drug network. We assume that drugs having common target proteins are more related to each other. In the TD network, the thickness of an edge between drugs denotes the number of common target proteins. Through this network, we can infer trends in pancreatic cancer research by target proteins.

As shown in Fig. 7, the TD network has only one major component with strong local clustering value and many subgroups. We colored drug nodes according to data sources (ClinicalTrial.gov and PubMed): blue nodes denote drugs from clinical data and red nodes denote data from PubMed; green nodes are drugs from both datasets.

Out of 6319 drugs in DrugBank (Law et al., 2014), 210 are related to pancreatic cancer. Most of the drugs contained in our data are located at the main hub of the TD network. Drugs appearing in data from both ClinicalTrials.gov and PubMed are positioned at the upper side of the network (A) and includes bosutinib, gemcitabine, doxorubicin, fluorouracil, capecitabine, and cisplatin. These are common drugs for chemotherapy that belong to the “Standard therapeutic agent” and “Target agent” clusters in clinical data and “Chemotherapy agent” and “Antitumor effect” in PubMed. All of these drugs have a target “Thymidylate synthase” and significantly benefited patients with high tumor expression of this compound (Giovannetti

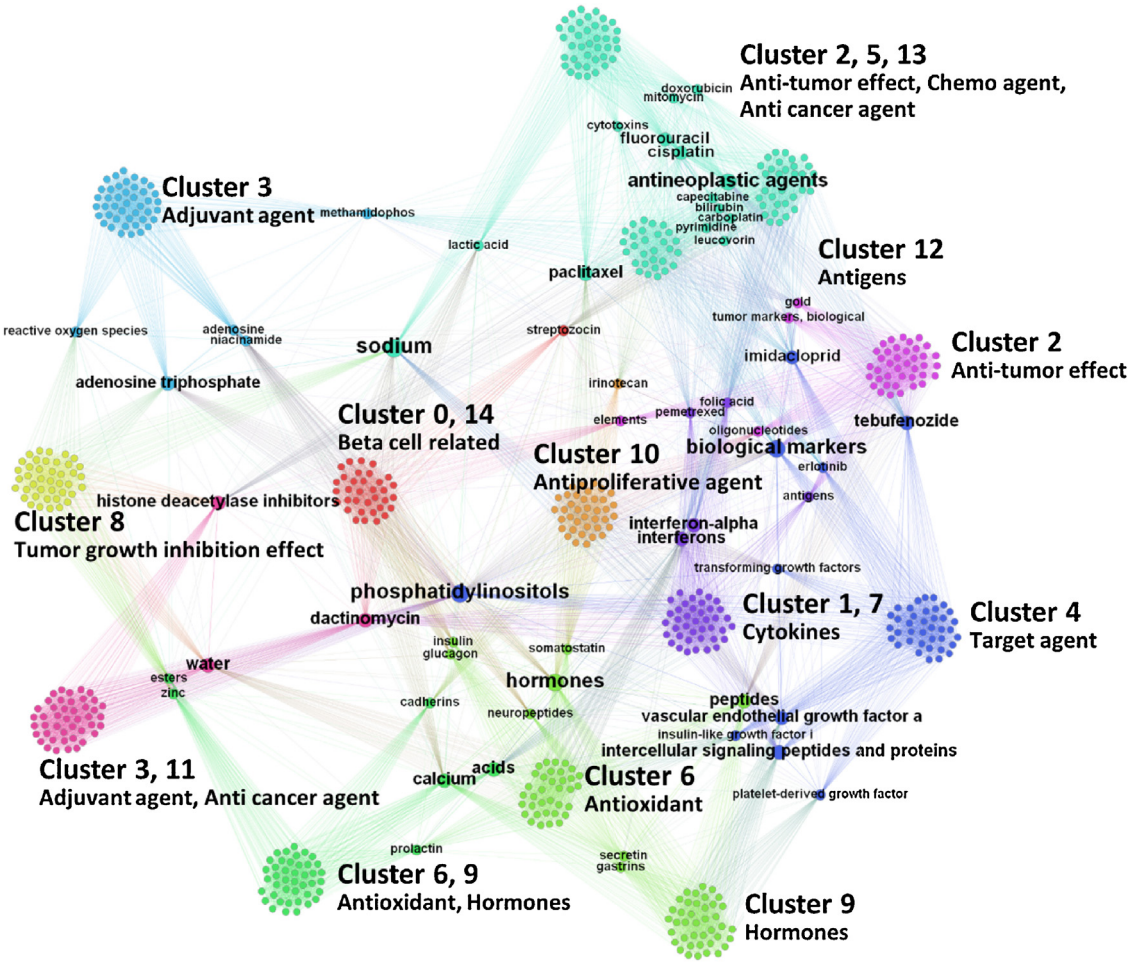

Fig. 6. Drug cluster network from PubMed records.

et al., 2006). The other targets like "Alpha-1A (or 2A) adrenergic receptor," "DNA topoisomerase 2-alpha," "Prostaglandin G/H synthase 2," and "Acetylcholinesterase" play a crucial role in pancreatic cancer-related drug research. These results indicate that the important drugs used for chemotherapy share similar proteins or enzymes as targets.

Drugs in clinical-data clusters that are on the rise, such as new-drug testing and modified anticancer drugs, are spread widely over the network (B, C, and D). In contrast, drugs in PubMed clusters showing a growing trend (target agent, chemotherapy agent, and tumor growth inhibitor effect) are lumped together (near C). This suggests that intensively researched drugs are linked via a particular set of targets, whereas drugs in clinical trials are tested more rationally with various targets. The results do, however, show various chemicals have been researched in PubMed, as compared to the limited targets in the TD network.

4. Discussion

We compared drug-research trends using a drug-cluster analysis and TD network analysis of records on PubMed and ClinicalTrials.gov. First, we analyzed the time gap in published records that mention drugs appearing in both databases, such as gemcitabine, fluorouracil, oxaliplatin, and capecitabine (Fig. 8). In the Fig. 8, the *Y*-axis is the number of publications and *X*-axis is publication year.

PubMed publications for gemcitabine and fluorouracil (upper graphs in Fig. 8) precede the mentions of these drugs in clinical data. In addition, PubMed publications include relevant research more frequently than the clinical data records do. In gemcitabine case, PubMed records first published in 1993 and the number of publications in the 2000s has been consistently rising in PubMed. However, clinical trials first tested the drug in 1999. Fluorouracil experienced about a 10-year gap between being mentioned in PubMed and on ClinicalTrials.gov. Research on this drug peaked in 2000, when a Phase 2 study of combination chemotherapy between fluorouracil, gemcitabine, and other chemicals was ongoing (Ahmed et al., 2000; Berlin et al, 2000; Burch et al., 2000; Ikeda et al., 1999; Kurtz et al., 1999; Matano et al., 2000; Oettle et al., 2000). The

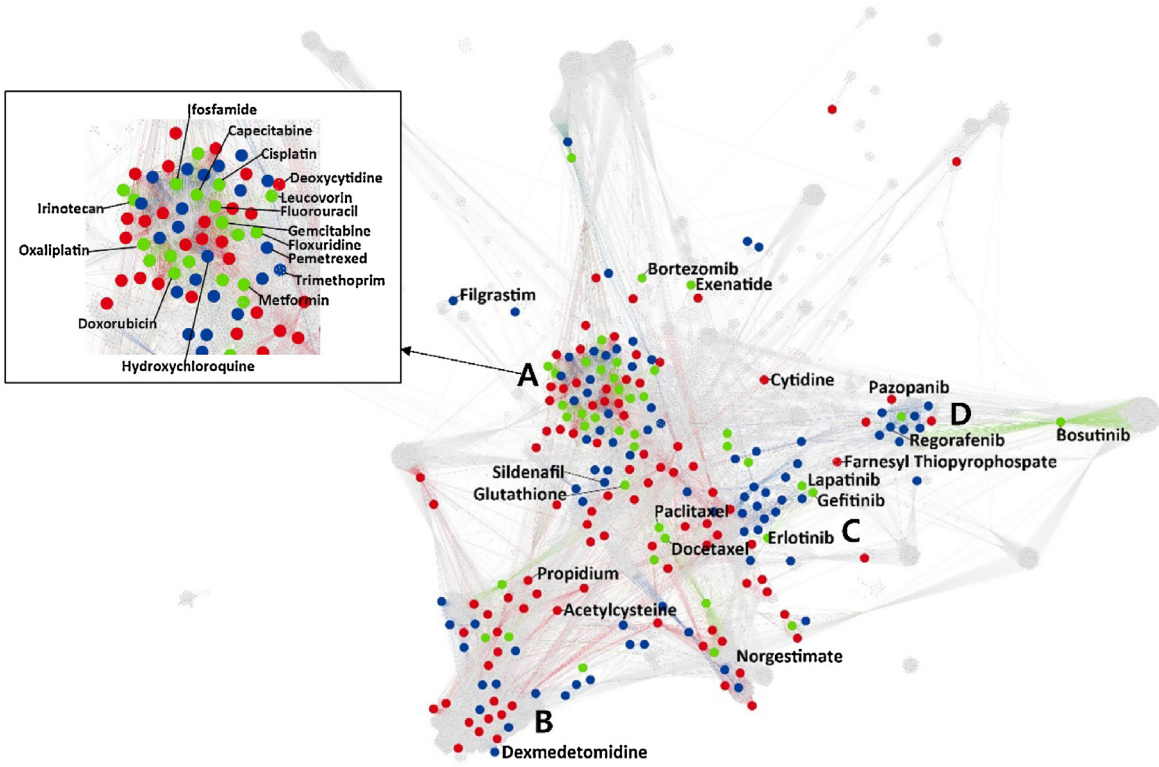

Fig. 7. Target-based drug network.

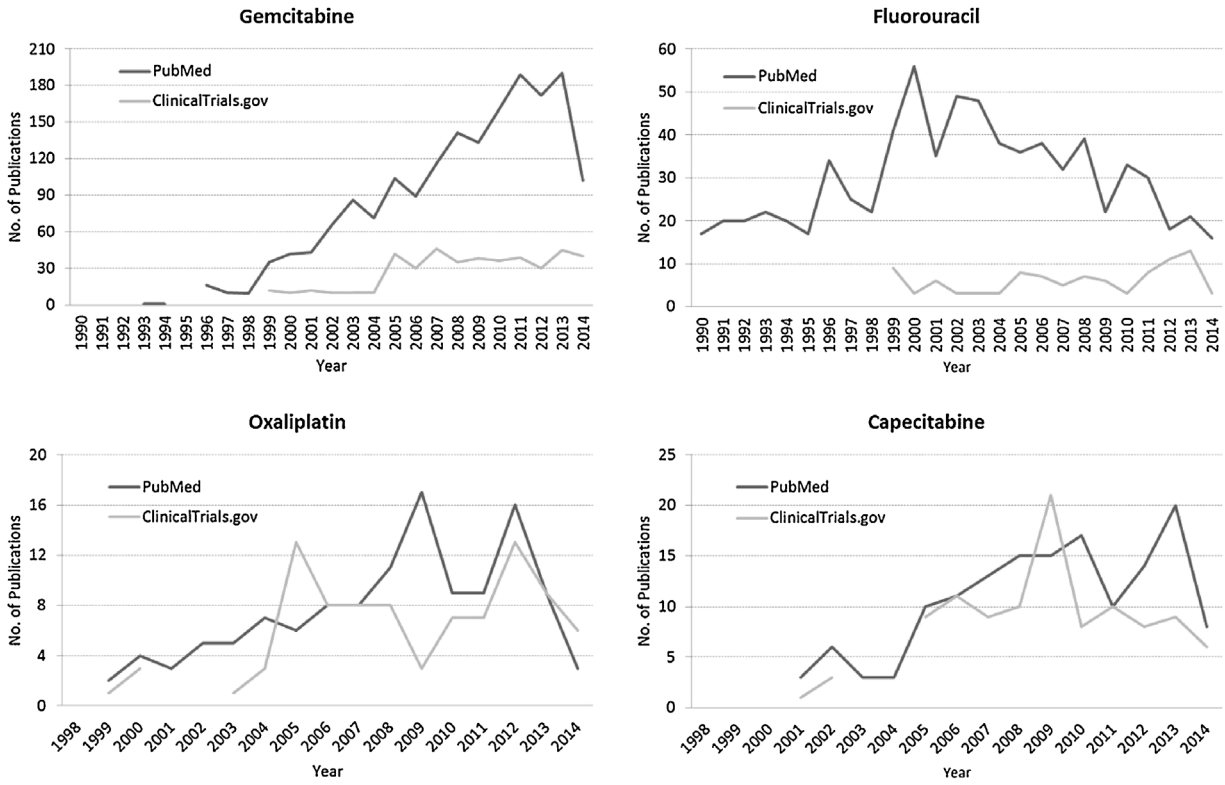

Fig. 8. Time gap in published records.

Table 6
Coverage of drug and chemical in Clinical data and PubMed.

	ClinicalTrial.gov	PubMed
Drug cluster		
# of drugs	810	3618
Clustering coefficient	0.980	0.948
TD network		
# of drugs	123	126
# of target proteins	503	473

results of these studies indicated that gemcitabine provides considerably higher clinical benefits than fluorouracil does. The combination of these two drugs was well tolerated and produced a symptomatic relief in the majority of tested patients. To evaluate the time lag between two data sets, we conducted a paired *t*-test using the 103 most-commonly studied drugs for more than five years. The results showed that PubMed research is ahead of ClinicalTrials.gov data by an average of 2.9 years with a large effect size (Cohen's $d = .88$).

There are exceptions, however. For the drugs oxaliplatin and capecitabine (lower side in Fig. 8), clinical trials precede publications on PubMed. Oxaliplatin is an antineoplastic agent that is combined with fluorouracil and folinic acid as the initial therapy for treatment of metastatic colorectal carcinoma and rectal cancer. It was tested in several clinical trials between early 2000 and 2008. Then in 2009, scholarly publications became common, while the number of trials decreased dramatically. Moreover, modified FOLFOX (a combination of folinic acid, fluorouracil, and oxaliplatin) regimens were tolerated with manageable toxicity (Yoo et al., 2009). However, in the case of cetuximab plus gemcitabine/oxaliplatin (GEMOXCET) (Kullmann et al., 2009) and oxaliplatin versus gemcitabine compared with gemcitabine research (Poplin et al., 2009), tested patients did not show improved response or survival over standard single-agent gemcitabine.

Another exceptional case is capecitabine (lower right side in Fig. 8). Capecitabine is mainly used for treatment of pancreatic cancer and colorectal cancer, and is enzymatically converted to fluorouracil as an inactive prodrug that is activated by the body's metabolic processes. Papers mentioning capecitabine were published most frequently studied on PubMed in 2013, while clinical trials were most common in 2009. In recent papers on PubMed, there have been various studies about combination therapies involving capecitabine (Boeck et al., 2013; Herman et al., 2013; Kordes et al., 2013; Lopez et al., 2013; Rajagopalan et al., 2013). Especially, erlotinib can be safely administered with chemotherapy by the result of the study (Herman et al., 2013). In the research of combination chemotherapy, a phase III trial has been conducted with more delicate method (Boeck et al., 2013) and adopted as more extensive application in recent years (Rajagopalan et al., 2013; Wo et al., 2014). Similar to oxaliplatin, the studies treating capecitabine also reported the results of clinical test as evidence of drug efficacy. Based on these observations, we can infer that some recently developed drugs are actively discussed and studied in PubMed publications after clinical tests were conducted.

There is another difference in the coverage of drugs between PubMed and clinical data (Table 6). The number of drugs mentioned on PubMed is about four times greater than the number of drugs studied in clinical trials. Likewise, the topological properties of the network–such as its clustering coefficient, a measure of how nodes in a graph tend to cluster (Watts and Strogatz, 1998)–indicate that the drugs published in papers on PubMed are more diverse than those of clinical trials. The smaller clustering coefficient in PubMed (0.948) supports that drugs are distributed across the network, whereas the drugs treated in clinical trials are densely connected. In other words, the clinical trials tend to focus on drugs with limited coverage, whereas the PubMed records encompass diverse pancreatic cancer drugs.

In contrast, the targets of drugs in clinical trials are more diverse than in PubMed, according to our TD network analysis. Although the number of drugs treated in each data set is similar (123 and 126 respectively), the number of targets in clinical trials is greater than the number in PubMed. This result is in line with the scattered pattern of trial drugs on the TD network (Fig. 7). In other words, clinical trials use more targets to discover new drug targets than PubMed research does.

5. Conclusion

Here we conduct trajectory analysis of the recent drug-research trends in pancreatic cancer on PubMed and ClinicalTrials.gov. This is an essential step in understanding a research area, tracking research history, and discovering new research hypotheses. Unlike other bibliometric studies in cancer research, we use PubMed records and clinical data from ClinicalTrials.gov to analyze the drug-research trends in terms of topic coverage and knowledge transfer. For drug-cluster analyses, we adopt the eDMR technique, a variation of LDA. We also conduct in-depth network analyses of drug clusters and target proteins.

We identify different drug-research patterns between two data sources. First, the assumption that drug research published in PubMed is preceded by clinical trials is statistically confirmed in our study. Second, we see a research trend of new drug testing with various targets in clinical data. On the other hand, we see that diverse chemicals (e.g., sodium, phosphatidylinositols, and calcium) together with standard therapeutic agents are studied in scientific publications in PubMed. We employ in-depth network analyses to discover trends of both academic research and clinical trials, which could inform the direction of future research.

However, our approach uses a simple TF-IDF weight algorithm to measure the similarity between drugs. In the biomedical domain, a word in given text may convey conceptual meaning as a bio-medical entity. Although the TF-IDF reflects both the importance of a term and its discriminative power, this measure may not capture the semantics of the term properly because the similarity calculated based on word count. Therefore, we need to consider conceptual similarity, which includes information about sematic relationships or associations among words or concepts and not only text similarity.

Although the range of drugs we examined in the present study is limited to pancreatic cancer, our method can be applied to analyze drug-research trends of other diseases. As a follow-up study, we plan to apply literature-based discovery techniques to identify new biomarkers or drugs in networks of pancreatic anticancer drugs constructed from PubMed records and ClinicalTrials.gov data sets.

Acknowledgments

This work was supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT, and Future Planning through the National Research Foundation and by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2012M3C4A7033342).

References

Aggarwal, B. B., Kumar, A., & Bharti, A. C. (2003). Anticancer potential of curcumin: Preclinical and clinical studies. *Anticancer Research*, *23*, 363–398.

Ahmed, S., Vaitkevicius, V. K., Zalupski, M. M., Du, W., Arlauskas, P., Gordon, C., et al. (2000). Cisplatin, cytarabine, caffeine, and continuously infused 5-fluorouracil (PACE) in the treatment of advanced pancreatic carcinoma: a phase II study. *American Journal of Clinical Oncology*, *23*(4), 420–424.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., et al. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, *13*(1), 161.

Berlin, J. D., Adak, S., Vaughn, D. J., Flinker, D., Blaszkowsky, L., Harris, J. E., et al. (2000). A phase II study of gemcitabine and 5-fluorouracil in metastatic pancreatic cancer: An Eastern Cooperative Oncology Group Study (E3296). *Oncology*, *58*(3), 215–218.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Boeck, S., Jung, A., Laubender, R. P., Neumann, J., Egg, R., Goritschan, C., et al. (2013). KRAS mutation status is not predictive for objective response to anti-EGFR treatment with erlotinib in patients with advanced pancreatic cancer. *Journal of Gastroenterology*, *48*(4), 544–548.

Burch, P. A., Ghosh, C., Schroeder, G., Allmer, C., Woodhouse, C. L., Goldberg, R. M., et al. (2000). Phase II evaluation of continuous-infusion 5-fluorouracil, leucovorin, mitomycin-C, and oral dipyridamole in advanced measurable pancreatic cancer: A North Central Cancer Treatment Group Trial. *American Journal of Clinical Oncology*, *23*(5), 534–537.

Conroy, T., Gavoille, C., Samalin, E., Ychou, M., & Ducreux, M. (2013). The role of the FOLFIRINOX regimen for advanced pancreatic cancer. *Current Oncology Report*, *15*(2), 182–189.

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., et al. (2012). The comparative toxicogenomics database: Update 2013. *Nucleic Acids Research*, *41*(D1), D1104–D1114.

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, *47*, 1–10.

Giovannetti, E., Del Tacca, M., Mey, V., Funel, N., Nannizzi, S., Ricci, S., et al. (2006). Transcription analysis of human equilibrative nucleoside transporter-1 predicts survival in pancreas cancer patients treated with gemcitabine. *Cancer Research*, *66*(7), 3928–3935.

Gullo, L., Pezzilli, R., & Morselli-Labate, A. M. (1994). Diabetes and the risk of pancreatic cancer. *New England Journal of Medicine*, *331*(2), 81–84.

Herman, J. M., Fan, K. Y., Wild, A. T., Hacker-Prietz, A., Wood, L. D., Blackford, A. L., et al. (2013). Phase 2 study of erlotinib combined with adjuvant chemoradiation and chemotherapy in patients with resectable pancreatic cancer. *International Journal of Radiation Oncology, Biology, Physics*, *86*(4), 678–685.

Huang, Q., Shen, H. M., Shui, G., Wenk, M. R., & Ong, C. N. (2006). Emodin inhibits tumor cell adhesion through disruption of the membrane lipid Raft-associated integrin signaling pathway. *Cancer Research*, *66*(11), 5807–5815.

Ikeda, M., Okada, S., Ueno, H., Okusaka, T., Tanaka, N., Kuriyama, H., et al. (1999). A phase II study of sequential methotrexate and 5-fluorouracil in metastatic pancreatic cancer. *Hepato-Gastroenterology*, *47*(33), 862–865.

Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., et al. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine*, *6*(7), 57.

Kordes, S., Richel, D. J., Klümpen, H. J., Weterman, M. J., Stevens, A. J., & Wilmink, J. W. (2013). A phase I/II, non-randomized, feasibility/safety and efficacy study of the combination of everolimus, cetuximab and capecitabine in patients with advanced pancreatic cancer. *Investigational New Drugs*, *31*(1), 85–91.

Kullmann, F., Hollerbach, S., Dollinger, M. M., Harder, J., Fuchs, M., Messmann, H., et al. (2009). Cetuximab plus gemcitabine/oxaliplatin (GEMOXCET) in first-line metastatic pancreatic cancer: A multicentre phase II study. *British Journal of Cancer*, *100*(7), 1032–1036.

Kurtz, J. E., Kohser, F., Negrier, S., Trillet-Lenoir, V., Walter, S., Limacher, J. M., et al. (1999). Gemcitabine and protracted 5-FU for advanced pancreatic cancer. A phase II study. *Hepato-Gastroenterology*, *47*(35), 1450–1453.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research*, *42*(D1), D1091–D1097.

Lewison, G., Purushotham, A., Mason, M., McVie, G., & Sullivan, R. (2010). Understanding the impact of public policy on cancer research: A bibliometric approach. *European Journal of Cancer*, *46*(5), 912–919.

López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2008). The actual citation impact of European oncological research. *European Journal of Cancer*, *44*(2), 228–236.

Lopez, R., Méndez, C. M., Fernández, M. J., Reinoso, C. R., Aldana, G. Q., Fernández, M. S., et al. (2013). Phase II trial of erlotinib plus capecitabine as first-line treatment for metastatic pancreatic cancer (XELTA study). *Anticancer Research*, *33*(2), 717–723.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55–60).

Matano, E., Tagliaferri, P., Libroia, A., Damiano, V., Fabbrocini, A., De Lorenzo, S., et al. (2000). Gemcitabine combined with continuous infusion 5-fluorouracil in advanced and symptomatic pancreatic cancer: A clinical benefit-oriented phase II study. *British Journal of Cancer*, *82*(11), 1772.

Mela, G. S., Cimmino, M. A., & Ugolini, D. (1999). Impact assessment of oncology research in the European Union. *European Journal of Cancer*, *35*(8), 1182–1186.

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. ar**Xiv preprint ar**Xiv:1206.3278. Available from ⟨ht**tp://mimno.infosci.cornell.edu/papers/dmr-uai.pdf⟩. URL: ⟨h**ttp://arxiv.org/abs/1206.3278⟩.

NCI, (2012). NCI Portfolio Analyses—Pancreatic Cancer: A Summary of NCI's FY2010 and FY2011 Portfolio and Selected Research Advances, 2012. Available from ⟨http://www.cancer.gov/types/pancreatic/research/pancreatic-research-progress-2012.pdf⟩. URL: ⟨http://www.cancer.gov/aboutnci/servingpeople/cancer-research-progress/portfolio-analyses⟩ (accessed 28.02.15).

Oettle, H., Arning, M., Pelzer, U., Arnold, D., Stroszczynski, C., Langrehr, J., et al. (2000). A phase II trial of gemcitabine in combination with 5-fluorouracil (24-hour) and folinic acid in patients with chemonaive advanced pancreatic cancer. *Annals of Oncology*, *11*(10), 1267–1272.

Paul, R., Ewing, C. M., Jarrard, D. F., & Isaacs, W. B. (1997). The cadherin cell–cell adhesion pathway in prostate cancer progression. *British Journal of Urology*, *79*(S1), 37–43.

PharmaCyte Biotechm, Inc, Pancreatic Cancer description. Available from ⟨http://www.nuvilex.com/pancreatic-cancer⟩ (accessed on 28.02.15).

Poplin, E., Feng, Y., Berlin, J., Rothenberg, M. L., Hochster, H., Mitchell, E., et al. (2009). Phase III, randomized study of gemcitabine and oxaliplatin versus gemcitabine (fixed-dose rate infusion) compared with gemcitabine (30-minute infusion) in patients with pancreatic carcinoma E6201: A trial of the Eastern Cooperative Oncology Group. *Journal of Clinical Oncology*, *27*(23), 3778–3785.

Rajagopalan, M. S., Heron, D. E., Wegner, R. E., Zeh, H. J., Bahary, N., Krasinskas, A. M., et al. (2013). Pathologic response with neoadjuvant chemotherapy and stereotactic body radiotherapy for borderline resectable and locally-advanced pancreatic cancer. *Radiation Oncology*, *8*(1), 254.

Song, M., Kim, W. C., Lee, D., Heo, G. E., & Kang, K. Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. *Journal of Biomedical Informatics*, *57*, 320–332.

Tanabe, L., Xie, N., Thom, L. H., Matten, W., & Wilbur, W. J. (2005). GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, *6*(Suppl 1), S3.

Ugolini, D., Casilli, C., & Mela, G. S. (2002). Assessing oncological productivity: Is one method sufficient? *European Journal of Cancer*, *38*(8), 1121–1125.

Ugolini, D., & Mela, G. S. (2003). Oncological research overview in the European Union. A 5-year survey. *European Journal of Cancer*, *39*(13), 1888–1894.

Wasserman, S., & Faust, K. (1994). . *Social network analysis: Methods and applications* (8) New York and Cambridge, ENG: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

Wo, J. Y., Mamon, H. J., Ferrone, C. R., Ryan, D. P., Blaszkowsky, L. S., Kwak, E. L., et al. (2014). Phase I study of neoadjuvant accelerated short course radiation therapy with photons and capecitabine for resectable pancreatic cancer. *Radiotherapy and Oncology*, *110*(1), 160–164.

Yıldırım, M. A., Goh, K. I., Cusick, M. E., Barabási, A. L., & Vidal, M. (2007). Drug-target network. *Nature Biotechnology*, *25*(10), 1119–1126.

Yoo, C., Hwang, J. Y., Kim, J. E., Kim, T. W., Lee, J. S., Park, D. H., et al. (2009). A randomised phase II study of modified FOLFIRI. 3 vs modified FOLFOX as second-line therapy in patients with gemcitabine-refractory advanced pancreatic cancer. *British Journal of Cancer*, *101*(10), 1658–1663.

Zhao, D., & Weng, C. (2011). Combining PubMed knowledge and EHR data to develop a weighted Bayesian network for pancreatic cancer prediction. *Journal of Biomedical Informatics*, *44*(5), 859–868.