텍스트마이닝을 이용한 약물유해반응 보고자료 분석

김현희¹ · 유기연^{2*}

¹동덕여자대학교 정보과학대학, ²동덕여자대학교 약학대학
(2017년 9월 15일 접수 · 2017년 12월 6일 수정 · 2017년 12월 6일 승인)

Analysis of Adverse Drug Reaction Reports using Text Mining

Hyon Hee Kim¹ and Kiyon Rhew^{2*}

¹College of Information Science, Dongduk Women's University, Seoul 02748, Republic of Korea

²College of Pharmacy, Dongduk Women's University, Seoul 02748, Republic of Korea

(Received September 15, 2017 · Revised December 6, 2017 · Accepted December 6, 2017)

ABSTRACT

Background: As personalized healthcare industry has attracted much attention, big data analysis of healthcare data is essential. Lots of healthcare data such as product labeling, biomedical literature and social media data are unstructured, extracting meaningful information from the unstructured text data are becoming important. In particular, text mining for adverse drug reactions (ADRs) reports is able to provide signal information to predict and detect adverse drug reactions. There has been no study on text analysis of expert opinion on Korea Adverse Event Reporting System (KAERS) databases in Korea. **Methods:** Expert opinion text of KAERS database provided by Korea Institute of Drug Safety & Risk Management (KIDS-KD) are analyzed. To understand the whole text, word frequency analysis are performed, and to look for important keywords from the text TF-IDF weight analysis are performed. Also, related keywords with the important keywords are presented by calculating correlation coefficient. **Results:** Among total 90,522 reports, 120 insulin ADR report and 858 tramadol ADR report were analyzed. The ADRs such as dizziness, headache, vomiting, dyspepsia, and shock were ranked in order in the insulin data, while the ADR symptoms such as vomiting, 어지러움, dizziness, dyspepsia and constipation were ranked in order in the tramadol data as the most frequently used keywords. **Conclusion:** Using text mining of the expert opinion in KIDS-KD, frequently mentioned ADRs and medications are easily recovered. Text mining in ADRs research is able to play an important role in detecting signal information and prediction of ADRs.

KEY WORDS: Data mining, drug-related side effects and adverse reactions adverse drug reaction, adverse drug reaction reporting systems, big data analytics

약물이상반응(adverse drug reaction, ADR)의 보고 시스템을 구축하고 세계적으로 데이터베이스화하여 ADR의 실마리 정보를 탐색하는 연구는 많은 국가 및 기관에서 다양한 통계적 불균형성 기반 분석(disproportionality analysis)방법을 활용하여 진행하고 있다. 예를 들면, 세계보건기구(World Health organization, WHO)의 WHO-Vigibase에서 활용하고 있는 Bayesian Confidence Propagation Neural network (BCPN) 분석방법,¹⁾ 유럽의약품청 (European Medicines Agency, EMA) Eudravigilance에서 활용하고 있는 proportional reporting ratio (PRR) 분석방법,²⁾ 네덜란드의 Lareb에서 활용하는 reporting

odds ratio (ROR) 분석방법³⁾ 등이 대표적인 실마리정보 탐색 방법으로 빈도계산자 분석을 통한 ADR 보고자료의 데이터베이스를 분석하여 실마리 정보를 탐색하는 방법이다.

그러나 최근 정보의 크기가 방대해지고, 그 형태가 다양해짐에 따라 새로운 분석방법을 활용한 ADR 보고자료를 분석하는 시도가 이루어지고 있다. 대표적인 분석방법으로 약물-약물 상호작용 연구에 활용된 연관규칙 분석과 ADR 인과관계 혹은 ADR 발생 예측을 위해 시도된 텍스트마이닝을 제시할 수 있다. 연관규칙 중 apriori 알고리즘을 활용하여 약물-약물 상호작용의 실마리를 찾고 이는 빈도계산자 방법론을 보조할

*Correspondence to: Ki Yon Rhew, College of Pharmacy, Dongduk Women's University, 60 Hwarang-ro 13-gil, Seongbuk-gu, Seoul 02748, Republic of Korea

Tel: +82-2-940-4159, Fax: +82-2-940-4159

E-mail: kiyon@dongduk.ac.kr

수 있다는 결론을 도출하였으며,^{4,5)} 텍스트마이닝은 트위터 등 소셜미디어에서 환자들은 자신이 복용한 약물과 자신에게 발생했을 것으로 예측하는 ADR을 작성하고 공유한다는 것에서 이러한 SNS 텍스트를 분석한 결과를 해석하였을 때, ADR을 모니터링하는 보조적인 방법으로 활용가능하다는 점을 시사하였다.^{6,7)} 또한 ADR의 SNS 데이터나 동일 질환 환자 온라인 모임 등의 텍스트 자료를 분석하여 인과관계를 분석하는 연구 등이 보고되었다.^{8,9)}

텍스트마이닝은 텍스트 분석을 위한 데이터마이닝 방법으로, 의학 분야에 비정형 데이터로부터 유용한 정보를 추출하는데 중요한 역할을 할 수 있다. 이는 기존의 데이터마이닝 분석 기법을 텍스트에 적용 가능하도록 변형하여 사용하며, 키워드들의 군집 및 분류를 가능하게 한다. 즉, 텍스트마이닝을 통하여 대량의 텍스트 문서로부터 주요 키워드를 추출할 수 있고, 특정 키워드와 함께 쓰인 연관 키워드를 통해서 키워드가 사용된 문맥을 유추할 수 있으며, 문서 내에서 유사한 의미로 사용된 키워드들을 그룹화하여 문서의 주제를 파악할 수 있다.

다른 한편으로는 개인 맞춤형 헬스케어 산업이 활성화됨에 따라 헬스케어 데이터를 활용한 의학 정보 서비스 제공에 대한 관심이 높아지고 있고, 실제로 환자 의무기록과 같은 정형 데이터 외에 많은 양을 차지하고 있는 의학 전문가의 소견, 소셜 미디어에 개인들이 올린 ADR 정보, 연구 결과 출판물 등¹⁰⁾ 여러 종류의 비정형 텍스트 데이터를 분석하고자 하는 노력을 기울이고 있다.^{11,12)}

텍스트마이닝 분석 방법은 대량의 텍스트 문서를 컴퓨터가 분석하여 일차적으로 중요한 키워드를 찾아내는 것을 목표로 한다. 이들 중요 키워드와 함께 사용된 연관 키워드를 통해서 문서의 문맥이나 핵심 내용을 파악할 수 있고, 키워드의 확률적 분포를 이용하여 문서를 분류하거나 주제를 찾을 수 있다.^{13,14)}

본 연구에서는 한국의약품안전관리원의 의약품이상사례보고시스템을 통해서 구축된 의약품부작용보고자료 (KAERS Database)로부터 전문가 의견 텍스트 정보만을 추출하여 텍스트마이닝을 시행하고자 한다. 기존의 일반인의 텍스트 데이터를 분석한 결과와는 다르게 의학 전문가 의견 텍스트는 전문가로서의 ADR 평가, 환자로부터 제공받은 체질적 특이사항, 치료 시 복용 중이던 다른 약물 및 과거 약물 알러지력, 기저질환 등을 포함하고 있다. 따라서 기존의 정형 데이터 분석과는 차별적으로 전문가의 구체적인 의견에 대한 정보를 통해 중요 키워드를 추출하고, 연관 키워드를 찾아내고, 유사 키워드를 군집화하여 파악하고, 이를 통해 구체적인 ADR 관련 내용뿐만 아니라 전문가 의견의 구조와 특성을 이해하는데 도움을 줄 수 있을 것이다.

연구 방법

연구 자료 및 대상

본 연구에서는 KAERS로 보고된 의약품 등 부작용보고원 자료를 분석 가능한 형태로 만든 의약품부작용보고원자료 (KIDS KAERS Database, KIDS-KD)중에서 1989년에서 2015년 6월까지 보고된 종합 의견 내 전문가의견 텍스트가 본 연구 대상이며, 전문가의견의 텍스트 정보를 가진 보고 건수는 총 90,552건였다.

데이터 전처리 및 데이터 탐색

본 연구 자료의 경우 영어와 한글이 혼합된 텍스트로서 한글 전처리와 영문 전처리 과정을 동시에 진행하였다. 한글의 경우 “트라마돌이”와 “트라마돌을”과 같이 조사가 달라서 다른 단어로 인식되는 일이 없도록 하기 위해서 조사를 삭제하는 작업을 수행하였다. 또한 분석에 필요한 주요 단어들이 모두 명사임에 주목하여 명사만을 추출하여 분석을 진행하였다. 영문의 경우 주로 등장하는 단어들이 약품명, 부작용명 등으로 and, or, not, a, the 등과 같은 의미가 없이 자주 사용되는 단어들을 우선적으로 삭제하였으며, 특수문자 및 구분자, 날짜 시간 등도 삭제하였다. 또한 원인 약물이나 증상과 상관없이 의미 없이 반복되는 중복 텍스트도 삭제하였다.

전처리된 텍스트 문서가 분석에 적절한지 파악하고 구체적으로 분석하고자 하는 의약품 찾기 위해서 데이터의 특성을 파악하는 단계를 데이터 탐색이라고 한다. 이를 위해 총 90,522개의 데이터 중에서 10%를 랜덤 샘플링하여 단어에 대한 빈도수 분석을 실시하였으며, 심각한 이상반응으로 보고된 자료에 포함된 상위 5개의 의약품(sorafenib, insulin, aspirin, other drug, tramadol) 중 insulin, tramadol로 제형 및 기존 문헌에서의 ADR의 심각도/빈도를 고려하여 선정하였다.

통계 분석

본 연구에서는 먼저 빈도수 분석을 통해서 전반적으로 문서에 많이 등장한 단어들을 살펴보고, 좀더 의미 있는 중요한 단어를 찾기 위해 TF-IDF(term frequency-inverse document frequency) 가중치를 계산하여 중요 단어를 추출하였다. TF-IDF 가중치는 다음 수식 (1)과 같이 계산된다.¹⁵⁾

$$w_{(i,j)} = tf_{(i,j)} \times \log\left(\frac{N}{df(i,j)}\right) \quad (1)$$

문서에 속하는 키워드 i 의 가중치 $w_{(i,j)}$ 는 키워드 i 의 빈도수에 보정값을 곱한 것이다. 보정값은 키워드 i 를 포함하는 문서 수의 역수에 로그를 취한 값으로, 보정값을 곱하여 줌으로써 ‘너무’, ‘의약품’, ‘부작용’ 등과 같이 빈도수는 높으나 일반적인 단어로써 구체적인 정보를 제공하지 않는 단어들을 배제할 수 있

다. $f_{(i,j)}$ 은 문서내의 용어의 사용빈도, 즉 문서(d)에서 용어(t)가 발생하는 횟수를 의미하며, $df_{(i,j)}$ 는 단어가 제공하는 정보의 양, 즉 용어가 모든 문서에서 공통적으로 많이 사용되는 것인지 또는 희귀하게 나타나는지 여부를 나타내는 척도이다.

독립적인 단어만을 고려하면 단어가 사용된 문맥을 파악할 수 없으므로 중요 단어에 대한 보다 구체적인 정보를 추출하기 위해서 연관 단어들을 검색하였다. 이를 위해 선정된 단어와 다른 단어 간의 유사도를 Pearson correlation coefficient값을 사용하여 구하였으며, 임계치 내의 단어들을 추출하여 중요 단어와 함께 사용된 연관 단어들을 찾아내었다. 자료 분석을 위해서 통계 분석 프로그래밍 언어인 R을 사용하였으며, 텍스트마이닝을 위한 패키지인 tm package와 KoNLP 패키지를 활용하였다.¹⁶⁾

연구 결과

빈도수 분석

본 연구의 목표데이터로 인슐린을 포함한 ADR 보고의 전문가의견 텍스트 120건, 트라마돌 858건이 추출되었다. 각각 목표데이터의 전문가의견의 빈도수 분석 시행결과로는 총 120건의 인슐린 목표 데이터에서 dizziness (39회), headache (32회), glargine (30회), vomiting (28회), pentamidine (27회), 저혈당 (24회), lispro (24회), glucose (24회), weight (22회), dyspepsia (22회), shock (20회)의 순으로 나타났으며, 빈도수 상위100개 단어는 Figure 1에서 워드 클라우드를 통해 확인할 수 있다.

총 858건의 트라마돌 목표 데이터에서는 vomiting (622회), inj (607회), tridol (553회), hcl (489회), 엑스자드 (472회), 어지러움

Fig. 2. The highest top 100 frequency keywords in tramadol data.

(439회), dizziness (432회), dyspepsia (420회), constipation (384회), acetaminophen (378회) 순으로 나타났으며, Figure 2는 빈도수 상위 100개 단어를 워드 클라우드를 통해 보여준다.

두 경우 모두에서 많이 등장한 단어들은 ADR 증상을 나타내는 단어임을 알 수 있으며, 인슐린 데이터에서는 dizziness가 트라마돌 데이터에서는 vomiting이 가장 많이 등장하였다. 그밖의 텍스트 분석 결과로는 인슐린 데이터에서는 저혈당증, 소화불량, shock, 체중 감소 등의 증상이 많이 등장한 반면, 트라마돌 데이터에서는 가려움, 두드러기, 어지러움, 호흡곤란 등의 단어가 자주 등장한 단어로 나타났다.

Fig. 1. The highest top 100 frequency keywords in insulin data.

TF-IDF 가중치 적용 빈도수 분석

인슐린 데이터에서 TF-IDF 가중치를 적용한 분석 결과에 따르면 가장 중요한 단어는 가중치 값이 27.63인 pentamidine으로 나타났으며 shock가 20.72로 2위로 나타났다. 애피드라, 암로피딘, cell, care, protirelin, tartrate, stiglipin, determir, aspirin, diarrhoeas, rantidine이 모두 13.81로 동등하게 중요한 단어로서 분석되었고, edema와 leg가 11.81로 뒤를 이었다 (Table 1, Online Supplement 1).

트라마돌 데이터에서는 엑스자드가 가중치 값 128.18로 가장 중요한 단어로 나타났으며 stevens johnson 증후군과 타그나가 61.23으로 다음으로 중요한 단어로 나타났다. syndrome이 57.13, flank가 38.99로 그 뒤를 이었으며 epidermal, necrolysis, mecrolysis가 34.99로 나타났다. 아스피린 30.99, back, 어지러움, peration, 부종 그리고 pancreation이 29.24였으며 마지막으로 fatigue가 27.76으로 나타났다 (Table 1, Online Supplement 1).

TF-IDF 가중치 분석 결과 인슐린과 트라마돌 모두에서 약

Table 1. The most important keywords using TF-IDF weight in the target data (Insulin, Tramadol)

Target data	Rank	Keyword	(TF-IDF Value)
Insulin	1	Pentamidine	(27.63)
	2	Shock	(20.72)
	3	Apidra (kr)	(13.81)
	3	amlodipine (kr)	(13.81)
	3	Cell	(13.81)
	3	Care	(13.81)
	3	Protirelin	(13.81)
	3	Tartrate	(13.81)
	3	Stiglipitin	(13.81)
	3	Determir	(13.81)
	3	Aspirin	(13.81)
	3	Diarrhoea	(13.81)
	3	Rantidine	(13.81)
	14	Edema	(11.81)
	14	Leg	(11.81)
Tramadol	1	Exjade (kr)	(128.18)
	2	Stevensjohnson	(61.23)
	2	Tagna (kr)	(61.23)
	4	Syndrome	(57.13)
	5	Flank	(38.99)
	6	Epidermal	(34.99)
	6	Necrolysis	(34.99)
	6	Mecrolysis	(34.99)
	9	aspirin (kr)	(30.99)
	10	Back	(29.24)
	10	dizziness (kr)	(29.24)
	10	Peration	(29.24)
	10	edema (kr)	(29.24)
	10	Pancreation	(29.24)
	15	Fatigue	(27.76)

(kr) : reported in Korea

물명이나 의약품명이 중요 단어로 등장하였으며, 특히 상위 15개의 중요 단어 중에 인슐린과 트라마돌 데이터 모두에서 아스피린이 중요 단어로 선정되었다. 그 외의 단어들은 부작용을 나타내는 단어들로서 인슐린 데이터에서는 shock와 leg가 트라마돌 데이터에서는 어지러움, 부종, fatigue등이 상위 15개 중요 단어에 속하였다(Table 1, Supplement 1).

연관 단어 분석

먼저 중요 단어를 인슐린과 트라마돌 데이터에서 각 10개씩 선정하였다. 선정 방법은 빈도수 분석과 TF-IDF 가중치 분석 결과 상위 10위에 속하는 단어들 중에서 상관 계수를 적용하였을 때, 0.2 이상의 값을 갖는 연관 단어가 존재하는 단어들로 선

정하였다. 인슐린 데이터의 중요 단어는 dizziness, headache, weight, vomiting, dyspepsia, pentamidine, protirelin, 암로피딘, sitagliptin, 그리고 aspirin으로 선정하였다. 트라마돌 데이터의 중요 단어는 vomiting, 어지러움, dizziness, dyspepsia, constipation, tridol, fentanyl, acetaminophen, serotonin, 그리고 stevens johnson 증후군으로 선정하였다.

인슐린 데이터에서 상관 계수 값이 1.0인 연관 단어를 갖는 중요 단어들은 pentamidine, protirelin, 암로피딘, 아스피린, 그리고 sitagliptin으로 대부분 의약품명이었으며, 연관 단어들은 주로 다른 의약품으로 이는 약물간 상호작용에 의한 부작용을 연구하는데 실마리 정보로 활용될 수 있다. 부작용 증상의 경우 상관 계수 값이 0.5 이상을 갖는 단어들을 살펴보면, headache와 decitabine tyleno이 0.63의 상관 관계를 보였고, weight와 decreased가 0.77, hyperglycaemia가 0.5의 상관 관계를 보였다(Table 2, Online Supplement 2).

트라마돌 데이터는 인슐린 데이터와는 달리 상관 계수가 1.0인 연관단어는 stevens johnson과 toxic 그리고 serotonin과 재흡수 억제 만으로 나타났다. Stevensjohnson 증후군의 경우 0.5이상의 상관 계수값을 갖는 단어들이 다수 등장하였음을 알 수 있다. 선정된 단어 중에서 vomiting, 어지러움, dizziness, dyspepsia, constipation과 같은 부작용과 함께 사용된 연관 단어들은 상관 계수가 0.3을 넘지 않는 것으로 나타났으며, 0.4 이상의 상관 관계를 보인 단어들은 fentanyl과 citrate, maxnophen이 0.42로 acetaminophen과 ultracet이 0.43으로 나타났다(Table 3, Online Supplement 3).

고찰 및 결론

KIDS에서는 ADR 보고가 많았던 다빈도 의약품이나 보고 자료에 포함된 환자의 기본사항(성별, 연령 등)을 정기적으로 분석하여 공개하고 있다. 그러나 전문가의견 등과 같은 텍스트 문서는 현재까지 분석에 활용되지 못하였기 때문에 이에 대한 아쉬움이 있었다.¹⁷⁾ 본 연구는 ADR 보고자료의 전문가의견 텍스트로 텍스트마이닝을 시행한 국내 첫 번째 연구결과로 그 의미가 크다고 할 수 있다. 이전 연구에서 환자의 텍스트 데이터를 분석하여 bisphosphonates 의약품이 심각한 근골격의 통증을 유발한다는 서한을 FDA로 보내고,¹⁸⁾ 몇 년 후 이러한 결과가 반영되어 조사가 착수된 사례처럼¹⁹⁾ 보고사례에 대한 전문가 의견은 전문가의 평가 소견 뿐 만 아니라 환자로부터 제공받은 정보들까지 구체적으로 기술되어 있어 장기간 추적적이 어려운 ADR이나 소아 및 노인 등 특정 환자의 ADR 실마리 정보를 파악하는데 유용하게 활용될 수 있을 것으로 생각된다.

목표데이터의 텍스트마이닝 결과 상위 빈도수를 차지하는 단어는 ADR과 관련된 증상 등을 나타내는 용어였으며, TF-

Table 2. Related keywords in the insulin data

Target Keyword	Related Keywords (correlation coefficient)
dizziness	Anorexia (0.3), ache (0.28), chills (0.28), diaphoresis (0.28), distension (0.28), fever (0.28), gaseous (0.28), hypoglycaemia (0.28), sulfonylurea (kr) (0.28), migraine (0.28), mood (0.28), phosphate (0.28), sitagliptin (0.28), sweat (0.28), nausea (kr) (0.28), dizziness (kr) (0.28), hypoglycemia (kr) (0.28)
headache	Decitabine Tylenol (0.63), Exjade (kr) (0.43), hyperglycaemia (0.34), fatigue (0.30), mood (0.30), nasopharyngitis (0.30), peridipine (0.30), preline (0.30), protirelin (0.30), tartrate (0.30)
weight	Decreased (0.77), hyperglycaemia (0.5), increase (0.44), parpitation (0.44), heartburn (kr) (0.44), heartburn (0.30)
vomiting	Abdominal (0.42), ache (0.37), distension (0.37), gaseous (0.37), glucagon (0.37), migraine (0.37), octreotide (0.37), polypeptide (0.37), secretin (0.37), somatostatin (0.37), sostatin (0.37), tooth (0.37), visipaque (0.37), intermittent gastrointestinal disorder (kr) (0.37), fluid intake (kr) (0.37), nausea (kr) (0.37)
dyspepsia	Increase (0.44), sedation (0.44), gastrointestinal (0.30), abdominal (0.23), hyperglycaemia (0.23)
pentamidine	Abnormality (1.0), carinii (1.0), isethionate (1.0), nephrotoxicity (1.0), pancreatic (1.0), pneumocystis (1.0), toxicity (1.0), hypoglycemia (1.0), cytolytic (1.0), azotemia (1.0), blood (1.0), metabolic (0.7) glucose (0.37)
protirelin	Perdipine (1.0), preline (1.0), inj (1.0), headache (0.3)
amlodipine (kr)	Glimep (1.0), mgbsa (1.0), pneumonia (kr) (1.0), heart function (kr) (1.0), body surface area (kr) (1.0), sepsis (kr) (1.0)
sitagliptin	Chills (1.0), fever (1.0), januvia (1.0), phosphate (1.0), sweating (1.0), fastic (1.0), chills (kr) (1.0), dizziness (kr) (1.0), hypoglycemia (kr) (1.0), glupa (0.89), hcl (0.70), dizziness (0.28)
aspirin	65 year old (1.0), aderamine (1.0), beraprost (1.0), biloba (1.0), carvedilol (1.0), ciprofloxacin (1.0), clozapine (1.0), concomitant (1.0), diarrhea (1.0), dietetic (1.0), dulcolax (1.0), Enderlin (1.0), erythropoietin (1.0), ginkgo (1.0), infection (1.0), irbesartan (1.0), ranitidine (1.0), recombinant (1.0), spironolactone (1.0), telmisartan (1.0), vitamedin (1.0), acid (0.7), gastrointestinal (0.7)

(kr) : reported in Korean

IDF 가중치를 적용한 분석에서는 주로 구체적인 약물명이 중요 단어로 나타났다. 그러나 전문가의견의 경우 보다 많은 정보를 기입하기 위해 구체적으로 작성하는 의료진과 의미있는 데이터를 거의 포함하지 않은 채 작성하는 의료진이 구분되어 있어 빈도수, TF-IDF 가중치 적용 빈도수, 연관규칙 분석에서 비뚤림이 나타났을 것으로 판단된다. 또한 오타로 추정되는 단어가 높은 빈도수를 차지하거나(예, 엑스자드, 타그나 등) 연관분석 시 상관관계가 1.0을 차지하는 결과가 도출되었던 것은 전문가가 의견을 모아서 작성할 때 동일 약물이나 증상에 대해 ‘복사하기 및 붙여넣기’를 반복적으로 시행함으로 나타난 결과라고 생각된다.

또한 전문가의견을 단순 빈도수 분석을 시행한 결과와 TF-IDF 가중치를 적용하여 빈도수 분석을 시행하였을 때 그 결과가 다름을 알 수 있고, 이는 TF-IDF 가중치를 활용하여 보다 단순 빈도분석 결과와는 차별적인 결과를 도출할 수 있을 것으로 판단된다. 예를 들면 인슐린의 경우 dizziness, headache, 저혈당, vomiting, dyspepsia, shock의 순으로 부작용이 언급되었으며, 트라마돌의 경우 vomiting, 어지러움, dizziness, dyspepsia, fentanyl의 순으로 부작용이 언급되었으나 TF-IDF 가중치 분석으로 인슐린 데이터에서 중요한 약물은 pentamidine, 애피드라, 암로피딘, protirelin, 아스피린 순으로 나타났다. 또한 트라마돌 데이터의 경우 stevensjohnson은 자주 언급되지 않은 단어로 분석되었지만 실제 가중치 값은 61.23로 두 번째로 높은 가중치 값으로 나타났고, 함께 사용된 단어들도 상관관계가 매

우 높게 분석되었다.

연관 분석결과를 살펴보았을 때는 빈도 분석과 TF-IDF 가중치 분석 결과 상위에 해당하는 단어들을 중요 단어로 선정하여 분석한 연관 단어들은 인슐린 데이터에서 pentamidine, protirelin, 암로피딘, aspirin등의 동시에 투약한 의약품으로 추정되는 의약품으로 실제로 상관계수가 1.0으로 나타나 동시에 투약한 약물을 전문가가 작성하였고 이 내용이 분석결과로 도출된 것으로 생각할 수 있다. 그러나 ADR 증상을 나타내는 단어는 실제 상관계수 값이 비교적 낮게 분석되어 환자마다 ADR 증상은 다양하게 나타나지만 복용하는 의약품이 비슷할 수 있다는 결과로 해석될 수 있다.

의약품 부작용 연구에서 텍스트 정보는 점차로 증가하고 있으므로 텍스트마이닝은 실마리 정보를 찾아내거나 부작용 정보를 예측하는데 중요한 역할을 할 것으로 기대된다. 본 연구에서는 보고자료 전체를 분석하였으므로, 인과관계 등 평가결과 정보를 적용하지 못했다는 제한점이 있다. 또한 텍스트마이닝 결과로 드러난 단어에서 구체적인 실마리정보를 탐색하기에는 보건의약 전문가들이 전문가의견 작성에 대한 필요성을 많이 느끼지 못하거나 현실적인 제약이 있을 수 있다. 그러므로 언급된 정보를 활용하여 부작용의 인과성 여부를 파악할 수 있는 추가적인 분석을 시행하거나 전문가의견을 작성하는 지침 등이 필요하다. 또한 전문가의견은 일반적으로 국문과 영문을 혼용하여 작성하는 경우가 많고, 텍스트마이닝의 기법이 어절 단위로 분석을 하기 때문에 두 개 이상의 어절이

Table 3. Related keywords in the tramadol data

Target Keyword	Related Keywords (correlation coefficient)
vomiting	Oxycodone (0.28), hydromorphone (0.25), lidocaine (0.25), pyridostigmine (0.25), acetaminophen (0.25), dizziness (kr) (0.25), xycodone (0.25)
dizziness (kr)	Pharbitis (0.37), tuber (0.37), Ultracet Tab. (kr) (0.37), Rhonal Tab. (kr) (0.37), Mobic Cap. (kr) (0.37), meloxicam (0.37), Motilitone Tab. (kr) (0.37), Exoperin Tab. (kr) (0.37), Enaprin Tab. (kr) (0.37), warfarin (0.37), anticoagulant (kr) (0.37) anti-inflammatory agent (kr) (0.32), airtal (0.31), aceclofenac (0.31), cetaminophen (0.31), motilitone (0.31)
dizziness	Ultrivaremfentanil (0.27), analgesic injection (kr) (0.27), headache (0.26), insomnia (0.20), joinsclamtis (0.20), mesocarnmesoglycan (0.20), mevalotinpravastatin (0.20), mshurica (0.20), eupatilin (0.20), eperisone (0.20), myonal (0.20)
dyspepsia	Abdominal (0.24), flatulence (0.24), isepamicin (0.21)
constipation	Xerostomia (0.39), oxycodone (0.29), durogesic (0.23), ircodon (0.23)
tridol	Inj (0.37), norepinephrine (0.33), resorption inhibition (kr) (0.33), analgesic action (kr) (0.33), receptor binding (kr) (0.31), saline (0.20), trolac (0.20), tylenol (0.20), ultracet (0.20), abdominal (0.20)
fentanyl	Citrate (0.42), maxnophen (0.42), fentanyl (0.36), maxnophen (0.31), headache (0.31), xerostomia (0.22), fluid (0.21), vomiting (0.21)
acetaminophen	Ultracet (0.43), maxnophen (0.39), hcl (0.36), aspirin (kr) (0.28), ultracet (0.28), analgesics (kr) (0.28), aspirin (0.27), allergy (kr) (0.27), tylenol (0.24), hypersensitivity (kr) (0.23), pruritus (kr) (0.22), opioid (0.20)
serotonin	resorption inhibition (kr) (1.0), norepinephrine (0.99), analgesic action (kr) (0.98), receptor binding (kr) (0.90), analgesic injection (kr) (0.27), nausea (kr) (0.23), hyperhidrosis (kr) (0.20)
stevensjohnson	Toxic (1.0), syndrome (0.99), cilostazol (0.76), sitagliptin (0.76), hydrochlorothiazide (0.76), mirtazapine (0.76), amithiozone (0.65), amoxicillin (0.65), antituberculosis (0.65), barbiturates (0.65), cotrimoxazole (0.65), diclofenac (0.65), ethambutol (0.65), fenbuten (0.65), fluoroquinolone (0.65), hydantoin (0.65), ketoprofen (0.65), lamotrigine (0.65), moxifloxacin (0.65), naproxen (0.65), nevirapine (0.65), opurinol (0.65), phenylbutazone (0.65), piroxicam (0.65), rifampin (0.65), sertraline (0.65), ssystem (0.65), sulfadiazine (0.65), sulfadoxine (0.65), sulindac (0.65), syndrometoxic (0.65), tenoxicam (0.65), thiabendazole (0.65), tiaprofenic (0.65), vancomycin (0.65)

(kr) : reported in Korean

하나의 의미를 가지는 경우에 이와 관련한 전처리 과정이나 텍스트마이닝 알고리즘을 그대로 적용하는 데에는 그 결과가 정교하게 도출되지 않은 점이 있다.

그럼에도 불구하고, 본 연구는 의미있는 목표 데이터를 선정 하여 이와 관련한 전문가의견을 텍스트마이닝을 시행함으로 실제로 전문가들이 ADR 증상이나 동시에 투여되는 약물을 빈도 높게 작성한다는 것을 알 수 있었고, 단어들 간의 연관분석을 통해 각각의 주요단어에서 어떤 단어를 의미 있게 제시하는 지를 보여줬던 것이 본 연구의 강점이라고 할 수 있다. 또한 국문과 영문이 혼용되어 있는 텍스트를 텍스트마이닝 분석 방법으로 결과를 제시한 것이 주요한 성과라고 할 수 있다.

결론적으로 KIDS-KD 자료내의 전문가의견을 텍스트마이닝하는 방법은 ADR 실마리 정보탐색이나 ADR 발생 환자의 특성을 파악하는데 추가적인 방법으로 시도 가능하며, 이는 기존의 실마리 정보탐색에 보조적인 역할을 담당할 수 있다. 또한 향후 인과 관계가 명확한 데이터나 개인별 입퇴원 정보나 복용 약물 정보 등만을 가지고 텍스트마이닝을 시행하는 등 다양한 방법론을 설계하여 보다 의미 있는 ADR 실마리 정보 탐색 방법을 제시할 수 있을 것으로 기대된다.

감사의 말씀

본 연구는 2016년도 식품의약품안전처의 연구개발비(2016 MFDS163)로 수행되었으며 이에 감사드립니다.

참고문헌

1. Bate A, Lindquist M, Edwards IR, *et al.* A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol* 1998;54:315-21.
2. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001;10:483-6.
3. van Puijenbroek EP, Bate A, Leufkens HG, *et al.* A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Saf* 2002;11:3-10.
4. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* 2010;11(suppl):S7.
5. Ibrahim H, Saad A, Abdo A, *et al.* Mining association patterns of drug-interactions using post marketing FDA's spontaneous reporting data. *J Biomed Inform* 2016;60:294-308.

6. Korkontzelos I, Nikfarjam A, Shardlow M, *et al.* Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *J Biomed Inform* 2016;62:148-58.
7. Duh MS, Cremieux P, Audenrode MV, *et al.* Can social media data lead to earlier detection of drug-related adverse events? *Pharmacoepidemiol Drug Saf* 2016;25:1425-33.
8. Harpaz R, Callahan A, Tamang S, *et al.* Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. *Drug Saf* 2014;37:777-90.
9. Wu L, Moh TS, Khuri N. Twitter Opinion Mining for Adverse Drug Reactions. In: *IEEE International Conference on Big Data*, Washington DC, USA, December 7, 2015.
10. Fang R, Pouyanfar S, Yang Y, *et al.* Computational Health Informatics in the Big Data Age: A Survey. *ACM Computing Surveys* 2016;49:12.
11. Raghupathi W, Raghupathi V. Big Data Analytics in Healthcare: Promise and Potential. *Health Inf Sci Syst* 2014;2:3.
12. Warrar P, Hansen EH, Juhl-Jensen L, *et al.* Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Br J Clin Pharmacol* 2011;73:674-84.
13. Kim HH, Kim D, Jo J. Patent Data Analysis using Clique Analysis in a Keyword Network. *Journal of the Korean Data and Information Science Society* 2016;27:1273-84.
14. Kim HH, Rhee HY. Trend Analysis of Data Mining Research Using Topic Network Analysis. *Journal of the Korea Society of Computer and Information* 2016;21:141-8.
15. Manning CD, Raghavan P, Schutze H. *Introduction to Information Retrieval*, 1st ed. New York; Cambridge University Press, 2008; 109, 116-21.
16. tm Package text Mining in R. Available from <https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>. Accessed June, 2017.
17. Botsis T, Nguyen MD, Woo EJ, *et al.* Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc* 2011;18:631-8.
18. Wysowski DK, Chang JT. Alendronate and risedronate: Reports of severe bone, joint, and muscle pain. *Arch Intern Med* 2005;165:346-7.
19. DeMonaco HJ. Patient- and physician-oriented web sites and drug surveillance: Bisphosphonates and severe bone, joint, and muscle pain. *Arch Intern Med* 2009;169:1164-6.