Journal of Health Informatics and Statistics

Original Article

J Health Info Stat 2018;43(4):329-335 • https://doi.org/10.21032/jhis.2018.43.4.329

pISSN 2465-8014 | eISSN 2465-8022

아토바스타틴의 새로운 약물 적응증 탐색을 위한 비정형 데이터 분석

정휘수[1], 강길원[2], 최 웅[3], 박종혁[2], 신광수[4], 서영성[5]

[1]동국대학교 의과대학 가정의학교실 교수, [2]충북대학교 의과대학 의료정보 및 관리학교실 교수, [3]충북대학교 의과대학 약리학 교실 교수, [4]충북대학교 의과대학 의생명과학경영융합대학원 교수, [5]계명대학교 의과대학 가정의학 교실 교수

Analysis of Unstructured Data on Detecting of New Drug Indication of Atorvastatin

Hwee-Soo Jeong[1], Gil-Won Kang[2], Woong Choi[3], Jong-Hyock Park[2], Kwang-Soo Shin[4], Young-Sung Suh[5]

[1]Professor, Department of Family Medicine, Dongguk University College of Medicine, Gyeongju; [2]Professor, Department of Health Information and Management, Chungbuk National University College of Medicine, Cheongju; [3]Professor, Department of Pharmacology, Chungbuk National University College of Medicine, Cheongju; [4]Professor, Graduate School of Health Science Business Convergence, Chungbuk National University College of Medicine, Cheongju; [5]Professor, Department of Family Medicine, Keimyung University School of Medicine, Daegu, Korea

Objectives: In recent years, there has been an increased need for a way to extract desired information from multiple medical literatures at once. This study was conducted to confirm the usefulness of unstructured data analysis using previously published medical literatures to search for new indications. **Methods:** The new indications were searched through text mining, network analysis, and topic modeling analysis using 5,057 articles of atorvastatin, a treatment for hyperlipidemia, from 1990 to 2017. **Results:** The extracted keywords was 273. In the frequency of text mining and network analysis, the existing indications of atorvastatin were extracted in top level. The novel indications by Term Frequency-Inverse Document Frequency (TF-IDF) were atrial fibrillation, heart failure, breast cancer, rheumatoid arthritis, combined hyperlipidemia, arrhythmias, multiple sclerosis, non-alcoholic fatty liver disease, contrast-induced acute kidney injury and prostate cancer. **Conclusions:** Unstructured data analysis for discovering new indications from massive medical literature is expected to be used in drug repositioning industries.

Key words: Text mining, Network analysis, Atorvastatin, Indication

서 론

매년 많은 의학연구가 출판되고 있는 요즘, 보건의료인들은 진료에 도움이 되는 새로운 정보를 얻기 위한 능력을 요구받고 있으며, 체계적 문헌고찰이나 메타 분석 등과 같이 연구결과들을 체계적으로 정리하여 결론 내는 방법들도 개발되었다[1,2]. 또한 임상문헌으로부터 원하는 지식을 추출하는 방법에서도 데이터 마이닝(data mining) 기법 등이 고려되었다[3].

최근에는 페이스북(facebook) 및 트위터(twitter) 등의 소셜 네트워크 서비스(social network service, SNS)에서 발생되는 비정형 빅데이터(big data) 분석을 위해 데이터 마이닝에 기초한 텍스트 마이닝(text mining)이나 소셜 네트워크 분석(network analysis) 등이 주목을 받는데[4], 이러한 비정형 빅데이터 자료 분석들은 임상에서 의약품 부작용 모니터링 자료 분석, 난치성 질환의 새로운 임상지표 탐색 및 신약 물질 발견 등에 활용되고 있다. 예를 들어 의약품 부작용 보고서를 텍스트 마이닝하여 중요 부작용 및 연관, 유사 부작용을 확인하거나[5], 연구문헌의 텍스트 마이닝을 통하여 알츠하이머병(Alzheimer disease)과 관련된 25개의 생화학적 지표 발견이 그 예이다[6]. 또한 신약 개발의 초기

Corresponding author: Gil Won Kang
1 Chungdae-ro, Seowon-gu, Cheongju 28644, Korea
Tel: +82-43-261-2838, E-mail: gilwon67@chunguk.ac.kr

Co-Corresponding author: Hwee-Soo Jeong
123 Dongdae-ro, Gyeongju 38066, Korea
Tel: +82-54-770-8286, E-mail: hweesoo@dongguk.ac.kr

Received: October 22, 2018 **Revised:** November 13, 2018 **Accepted:** November 19, 2018

No potential conflict of interest relevant to this article was reported.

How to cite this article:
Jeong HS, Kang GW, Choi W, Park JH, Shin KS, Suh YS. Analysis of unstructured data on detecting of new drug indication of atorvastatin. J Health Info Stat 2018;43(4):329-335. Doi: https://doi.org/10.21032/jhis.2018.43.4.329

It is identical to the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) whichpermit sunrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2018 Journal of Health Informatics and Statistics

단계에서 중국 전통의학 서적을 텍스트 마이닝하여 페룰산(Ferulic acid)이 주성분인 심황(Curcumin)이 알츠하이머병에 효과적임을 확인하여 향후 새로운 신약 물질로 보고되기도 하였다[7].

따라서 본 연구도 의학논문정보를 이용하여 기존의 의약품에서 새로운 적응증을 탐색하기 위하여 텍스트 마이닝 및 네트워크 분석을 수행하였다. 특히 본 연구에서는 심뇌혈관 질환의 주요 위험인자인 고지혈증 치료제인 하이드록시 메틸글루타릴 조효소A (Hydroxy-methylglutaryl Coenzyme A, HMG-CoA)환원효소 억제제들 중 아토바스타틴(Atorvastatin; Lipitor®, Pfizer, Inc., New York, USA)의 의학논문정보를 이용하였는데 HMG-CoA 환원효소 억제제 제품 중 아토바스타틴이 국내를 비롯하여 전 세계적으로 많이 처방되고 있으며 2008년에는 당뇨병 발생의 부작용이 보고된 이후 장기간 사용에 대해 논란 중이기 때문이다[8-10].

연구 방법

자료 수집 방법

2018년 6월 13일 저자들은 펍메드(PubMed)를 이용하여 메드라인(Medline)에 있는 아토바스타틴 연구 논문들을 추출하였다. 아토바스타틴의 MeSH (medical subject headings) 용어는 "Atorvastatin Calcium"이며 1990년에 등재되었으므로 검색기간을 1990년 1월 1일부터 2017년 12월 31일까지 설정하였다. 또한 논문들 중 초록이 이용 가능한 인간 대상 연구로 제한하였다. 최종 검색식은 "(Atorvastatin Calcium [MeSH Terms] OR Atorvastatin Calcium [All field] OR Atorvastatin [All field] OR Lipitor [All field] OR Liptonorm [All field]) AND has Abstract [text] AND ("1990/01/01" [PDat] : "2017/12/31" [PDat]) AND Humans [Mesh]"이었다. 이를 통하여 5,059편의 연구초록이 추출되었고, 이들을 엑셀(excel) 파일로 변환하였으며 초록이 누락된 2편을 제외하고 최종 5,057편의 연구초록을 이용하여 수행하였다.

분석 방법

텍스트 마이닝

텍스트 마이닝이란 대량의 텍스트 문서로부터 주요 키워드를 추출하여 키워드를 그룹화 혹은 문서의 주제를 파악하는 방법이다[5]. 본 연구는 5,057편의 아토바스타틴 연구초록을 이용하여 넷마이너 4.0 (Netminer, Cyram Inc., Seoul, Korea) 프로그램을 이용하여 분석하였다.

수집된 초록들에서 질병 키워드 추출을 위하여 명사를 제외한 동사, 대명사, 형용사, 부사, 수사 및 외국어 등의 키워드는 불용어로 선정한 후 1차 텍스트 마이닝을 하여 총 21,600개의 명사형 키워드를 추출하였다. 추출된 키워드들을 이용하여 저자들은 2차 키워드 추출을 위한 지정어 사전, 유사어 사전 및 제외어 사전 파일을 만들었다. 지정어 사전은 "Alzheimer", "Alzheimers", "AD"로 추출된 키워드들이 분리 없이 "Alzheimer Disease" 추출되도록 하기 위함이며, 유사어 사전은 대표어인 "Atorvastatin"이 "atorvastatin", "Atorvastatin", "Ator", "ATV", "ATC", "ATOR", "atorvostatin" 등 다양하게 초록에 표기된 것을 통일하기 위함이다. 제외어 사전에는 초록 양식에 사용되는 명사들(예: 'Background', 'Objective', 'Method' 등)이나 질병을 나타내지 않는 일반적 개념의 명사(예: 'Patient', 'Group' 등), 통계용어, 이름이나 단체를 나타내는 고유명사, 단위명사, 시간명사 등이 2차 키워드 추출 시 제외되도록 설정하였다. 이렇게 지정어, 유사어, 제외어 사전 파일을 이용하여 2단계의 텍스트 마이닝 결과에서는 최종 273개의 질병 키워드들이 추출되었다.

네트워크 분석

1930년대부터 사회과학분야에서 발달한 네트워크 분석은 사람관계 연구에서 대표적인 연구방법으로 자리를 잡은 후 자연과학, 의학, 경제 및 정책분야 등으로 다양하게 응용되고 있으며, 인터넷과 스마트폰 보급에 의한 소셜 미디어의 비정형자료 분석에 텍스트 마이닝과 결합되면서 텍스트 네트워크 분석(text network analysis) 혹은 언어 네트워크 분석(semantic network analysis)으로 진화가 이루어졌다[11]. 네트워크의 기본 개념은 노드(Node, 점)와 링크(Link)로 이루어진 구조를 말하며 노드는 사람이름이나 키워드와 같이 고유 식별이 가능하며 속성을 가진 정보를 말하고 링크는 이들 노드들의 관계를 의미한다. 네트워크 분석은 네트워크 내의 밀도(density), 연결정도(degree), 평균 거리(distance) 등의 평균값들에 대한 일반적 특성 확인과 네트워크 내 집중되는 특정 노드 추출(중심성 분석) 및 하위 구조에 대한 분석(하위그룹 분석)으로 이루어진다[12].

본 연구에서는 새로운 적응증 탐색이 목적이므로 출현빈도가 낮은 키워드를 제거하는 키워드 필터링 절차 없이 273개의 질병 키워드를 이용하여 네트워크 분석을 시행하였다. 분석을 위하여 전체 자료의 형태를 '논문×키워드' 형태의 이원 모드 네트워크(2-mode network)에서 공출현(Co-occurrence) 값에 기초한 '키워드×키워드' 행렬인 일원 모드 네트워크(1-mode network)로 자료를 변환하였다. 분석을 통하여 키워드들의 출현 빈도(frequency), 키워드 간의 링크 개수(number of link), 연결 관계, 밀도, 평균거리 등의 일반적인 네트워크 특성을 확인하였고, 네트워크에서 중심이 되는 키워드들을 추출하는 중심성(연결정도, 근접성, 매개성) 분석을 시행하여 각각 상위 10위 키워드들을 제시하였다. 또한 네트워크 내에 하위그룹 확인을 위하여 모듈성(modularity)을 이용 커뮤니티 분석을 시행하였다.

Table 1. Number and percentage of articles on atorvastatin by year

Year	1990	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
n	1	4	9	20	41	47	99	129	158	238	255	334
%	0.02	0.08	0.18	0.40	0.81	0.93	1.96	2.55	3.12	4.70	5.04	6.60
Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
n	317	303	359	313	338	347	317	300	323	304	284	217
%	6.27	5.99	7.10	6.20	6.68	6.86	6.27	5.93	6.39	6.01	5.62	4.29

단어 가중치법을 이용한 텍스트 마이닝

추출된 키워드들 중 기존의 적응증 키워드들을 배제하고 아토바스타틴의 새로운 질병 키워드 탐색을 위하여 단어 가중치 기법을 이용하여 텍스트 마이닝을 시행하였다. 단어 가중치 기법(Term Frequency-inverse Document Frequency, TF-IDF)은 단어의 단순빈도에 의존한 중요성 평가를 보완하고 문서들에서 흔히 출현하는 단어를 배제하여 특정단어를 추출하거나 문서들의 유사도를 결정하는 방법이다[13]. 따라서 아토바스타틴의 주요 적응증이면서 단순 출현 빈도가 높은 고지혈증, 고콜레스테롤혈증, 이상지질혈증과 주요 관련 질환인 심뇌혈관질환과 고혈압, 당뇨병에서의 TF-IDF 값을 확인하였고 이들을 제외하기 위한 TF-IDF 최종 절단점은 1.1을 이용하였다.

연구 결과

연구초록들의 일반적 특성 및 텍스트 마이닝 결과

총 5,057개의 아토바스타틴 연구 논문 초록들은 1990년 출판을 시작으로 2005년부터 2015년까지 매년 300편 이상이 발표되었으며 2017년에는 217편이 보고되었다(Table 1). 초록들에서 추출된 273개의 질병 키워드들 중 총 2,008번 출현한 당뇨병(Diabetes mellitus)을 포함하여 출현 빈도 상위 10위인 질병 키워드들로는 심혈관 질환(Cardiovascular disease), 관상동맥질환(Coronary artery disease), 급성 관상동맥 증후군(Acute coronary syndrome), 고콜레스테롤혈증(Hypercholesterolemia), 뇌졸중(Stroke), 심근경색증(Myocardial infarction), 이상지질혈증(Dyslipidemia), 죽상경화증(Atherosclerosis), 고혈압(Hypertension)이었다.

출현빈도 상위 50위인 키워드들로 구성된 네트워크 분석 결과

273개의 키워드들 중 출현빈도 상위 50위인 키워드들로 구성된 네트워크 분석 결과 밀도는 0.09 (범위 0-1)이었다. 키워드 간의 평균 연결관계는 2.34로 한 키워드가 2-3개의 키워드와 연결된 구조였다. 키워드 간의 평균 거리는 2.36로 키워드들이 2.4단계만에 연결되어 있었다. 키워드 간의 동시출현 확률을 의미하는 링크 값은 1,162개로 확인되었으며 동시출현 확률이 가장 낮은 0-0.05 구간 링크 값의 빈도가 전체 링크 값의 80%를 차지하였다. Figure 1은 출현빈도 상위 50위인 질병 키워드들을 이용하여 링크 값의 절단점을 0.05 이상으로 했을 때 네트워크를 시각화하였다.

Table 2. Top 10 keywords by centrality analysis and TF-IDF degree

No	Keywords by degree centrality	Score	Keywords by closeness centrality	Score
1	Diabetes mellitus	0.346	Diabetes mellitus	0.438
2	CVD	0.306	Dyslipidemia	0.423
3	Dyslipidemia	0.286	CVD	0.423
4	Atherosclerosis	0.285	CAD	0.416
5	Hypertension	0.265	Atherosclerosis	0.409
6	CAD	0.265	Hypertension	0.403
7	Myocardial infarction	0.225	Hypercholesterolemia	0.384
8	MACE	0.224	Myocardial infarction	0.373
9	ACS	0.204	MACE	0.357
10	Metabolic syndrome	0.183	Hyperlipidemia	0.352
No	Keywords by betweens centrality	Score	Keyword by TF-IDF	Score
1	Atherosclerosis	0.109	Atrial fibrillation	42
2	CAD	0.103	Heart failure	38
3	Dyslipidemia	0.095	Breast cancer	37
4	Diabetes mellitus	0.067	Rheumatoid arthritis	37
5	CVD	0.062	Combined hyperlipidemia	36
6	Depression	0.056	Arrhythmias	35
7	Hypertension	0.050	Multiple sclerosis	31
8	Hypertriglyceridemia	0.030	NAFLD	30
9	CKD	0.028	Contrast-induced AKI	27
10	Alzheimer Disease	0.028	Prostate cancer	27

ACS, Acute Coronary Syndrome; AKI, Acute Kidney Injury; CAD, Coronary artery disease; CKD, Chronic Kidney Disease; CVD, Cardiovascular disease; MACE, Major adverse cardiac events; NAFLD, Non-alcoholic fatty liver disease; TF-IDF, Term Frequency-Inverse Document Frequency.

중심성 분석에서는 한 키워드가 다른 키워드들과 연결된 정도를 의미하는 연결정도 중심성 값의 평균은 0.096±0.097 (범위 0-0.347)이었으며, 중심성 척도(centrality index)는 26.2%이었다. 연결정도가 높은 상위 10위의 키워드들은 당뇨병, 심혈관 질환, 이상지질혈증, 죽상경화증, 고혈압, 관상동맥질환, 심근경색증, 심장사건 후 주요부작용(major adverse cardiac events), 급성 관상동맥 증후군, 대사증후군(Metabolic syndrome)이었다. 다른 키워드들과 가까운 거리 여부를 확인하는 근

접 중심성 값의 평균은 0.231±0.146 (범위 0-0.439)이었고 중심성 척도는 24.7%이었다. 근접 중심성 값이 높은 상위 10개의 질병 키워드들은 당뇨병, 이상지질혈증, 심혈관 질환, 관상동맥질환, 죽상경화증, 고혈압, 고콜레스테롤혈증, 심근경색증, 심장사건 후 주요부작용, 고지혈증 순이었다. 키워드 간의 매개체 역할을 확인하는 매개 중심성 값의 평균은 0.015±0.028 (범위 0-0.109)이며 중심성 척도는 9.6%이었다. 매개 중심성 상위 10개의 질병 키워드들은 죽상경화증, 급성 관상동맥 질환, 이상지질혈증, 당뇨병, 심혈관 질환, 우울증, 고혈압, 고중성지방혈증(Hypertriglyceridemia), 만성신장질환(Chronic kidney disease), 알츠하이머병(Alzheimer disease)이었다(Table 2).

하위구조 파악을 위한 커뮤니티 분석에서는 모듈성 최적 절단점은 0.333점이었으며 이를 적용하여 네트워크를 하위구조로 분류 시 총 17개의 하위그룹으로 분류되었다. 이들 중 2개 이상의 키워드들이 포함된 하위그룹은 총 5개이었다(Figure 1). 가장 큰 그룹은 13개의 키워드로 구성된 그룹 17로 아토바스타틴의 주요 적응증인 이상지질혈증, 고콜레스테롤혈증, 고지혈증, 고중성지방혈증, 복합 고지혈증(Combined hyperlipidemia), 가족성 복합 고지혈증(Familial combined hyperlipidemia), 대사증후군, 자가면역 근염(Autoimmune myositis), 신증후군(Nephrotic syndrome), 비알콜성 지방간염(Non-alcoholic fatty liver disease), 비만(Obesity), 췌장염(Pancreatitis) 및 류마티스 관절염(Rheumatoid arthritis)으로 구성되었다. 두 번째는 10개의 키워드들로 구성된 그룹 15로 심혈관 질환, 관상동맥질환, 심근경색증, 급성 관상동맥 증후군, 죽상경화증, 심장사건 후 주요부작용, 협심증(Angina pectoris), 죽상동맥경화성 플라크(Atherosclerotic plaque), 뇌졸중(Stroke) 및 일과성 허혈 발작(Transient ischemic attack) 등 심뇌혈관 질환과 관련된 키워드들로 구성되었다. 세 번째로는 5개의 키워드로 구성된 그룹 16으

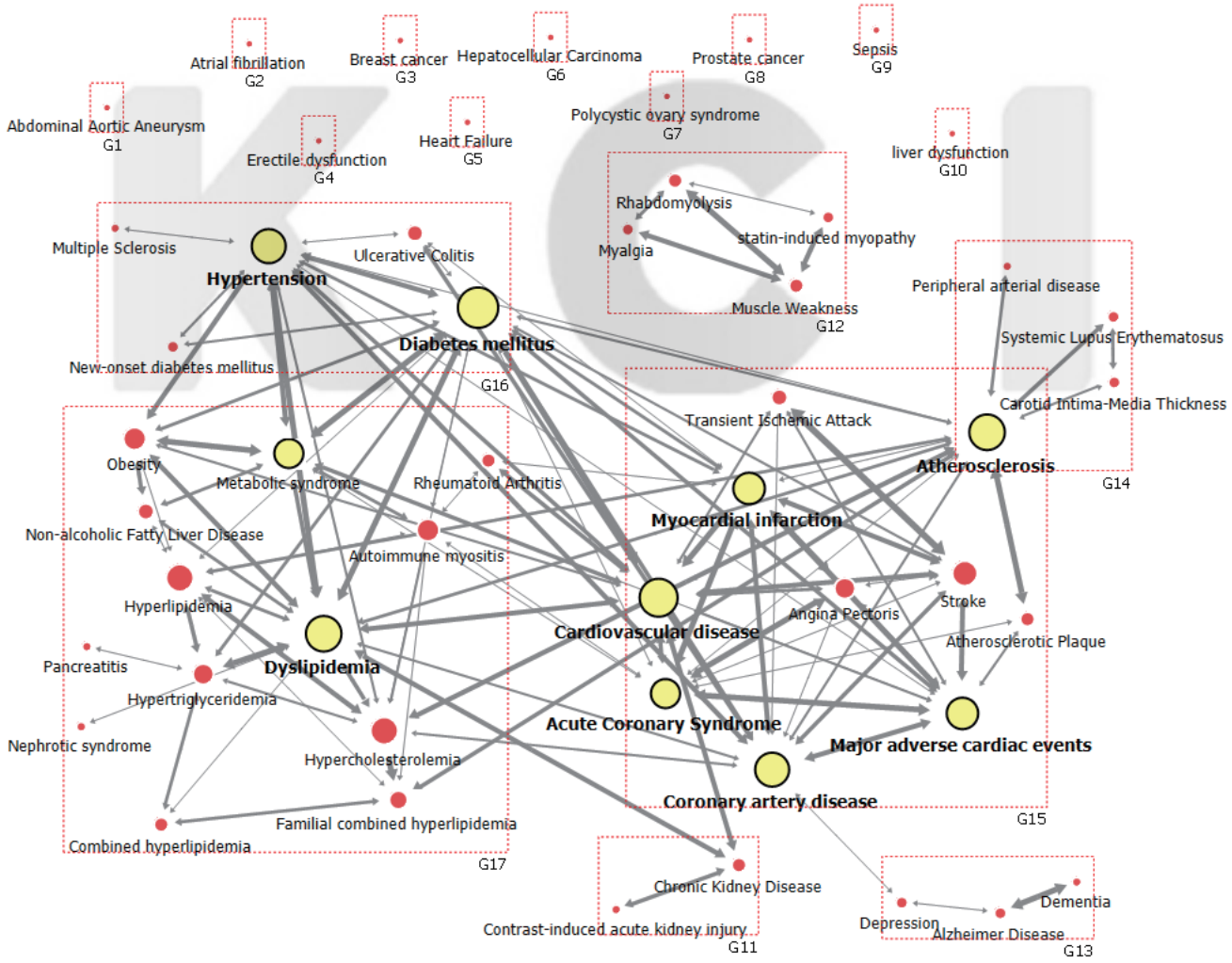

Figure 1. Whole network with Top 50 keywords by appearance. Frequency and a link value of 0.05 or higher. The higher score of degree centrality of keywords, the bigger size of node. The top 10 keywords of degree centrality of keywords were expressed as yellow node.

로 고혈압, 당뇨병, 새로 발병한 당뇨병(New onset diabetes mellitus), 다발성 경화증(Multiple sclerosis) 및 궤양성 대장염(Ulcerative colitis)으로 구성되었다. 네 번째는 네 개의 키워드로 구성된 그룹 12와 그룹 14로, 그룹 12에는 아토바스타틴의 부작용 키워드인 횡문근융해증(Rhabdomyolysis), 스타틴 유발 근염(Statin-induced myopathy), 근육통(Myalgia) 및 근무력(Muscle weakness)으로 구성되었으며, 그룹 14에는 죽상경화증과 관련된 말초 동맥 질환, 경동맥 내중막 두께(Carotid intima-media thickness), 전신성 홍반성 루프스(Systemic lupus erythematosus)로 구성되었다. 나머지는 알츠하이머병, 우울증(Depression), 치매(Dementia)로 구성된 그룹 13과 만성신장질환과 조영제 유발 급성 신손상(Contrast induced acute kidney injury)으로 구성된 그룹 11이다.

새로운 적응증 탐색을 위한 TF-IDF를 이용한 텍스트 마이닝

아토바스타틴의 새로운 적응증 탐색을 위하여 TF-IDF 절단점 1.1을 사용한 결과 273개의 키워드는 256개로 감소되었다. 이들 중 TF-IDF 정도값(degree)의 상위 10위인 키워드들은 심방 세동(Atrial fibrillation), 심부전(Heart failure), 유방암(Breast cancer), 류마티스 관절염, 복합 고지혈증, 부정맥(Arrhythmias), 다발성 경화증(Multiple sclerosis), 비알콜성 지방간염, 조영제 유발 신손상, 전립선암(Prostate cancer)이었다(Table 2).

고찰 및 결론

본 연구는 비정형 데이터 분석 도구인 텍스트 마이닝과 네트워크 분석을 이용하여 기존 약물의 연구초록에서 새로운 적응증 탐색을 위해 수행되었다. 총 5,057편의 아토바스타틴 연구초록에서 텍스트 마이닝을 통해 추출된 질병 키워드들은 273개이었으며, 이들 중 출현빈도 상위 50위인 키워드들로 네트워크 분석 결과 네트워크는 방향성이 없는 이진 네트워크로 하나의 키워드가 평균 2-3개의 키워드들과 연결되어 있고, 밀도는 0.09으로 높지 않아 키워드 간 관계가 유기적이라 할 수는 없었다.

중심성 분석들에서 나타난 상위 키워드들은 대부분 기존의 아토바스타틴 적응증인 고콜레스테롤혈증, 고지혈증, 이상지질혈증, 관상동맥질환, 급성 관상동맥 증후군, 죽상경화증, 심혈관 질환, 심근경색증, 고중성지방혈증, 복합 고지혈증으로, 화이자(Pfizer) 홈페이지에 소개된 아토바스타틴 적응증인 일차성 고지혈증, 이차성 이상지질혈증, 가족성 고콜레스테롤혈증(Familial hypercholesterolemia) 및 심근경색증, 뇌졸중, 협심증, 비치명적 심근경색증과 일치하는 키워드들이었다[14]. 이는 추출된 키워드들은 아토바스타틴 연구초록 내에서 필연적으로 등장하기 때문으로 생각된다. 또한 영향력 있는 키워드를 중심으로 군집화를 확인하기 위한 하위집단 분석에서도 심뇌혈관 질환 키워드들로 이루어진 집단을 중심으로 이들 질환의 위험인자인 이상지질혈증, 죽상경화증, 고혈압 및 당뇨병에 대한 하위집단들이 연결되어 있고, 횡문근융해증과 같이 약물의 부작용을 나타내는 키워드들로 별도의 그룹화됨으로써 네트워크 분석을 통하여 약물의 기존의 적응증과 부작용 키워드들이 연구초록에서 중요 키워드임을 확인할 수 있었다.

새로운 적응증 탐색을 위하여 TF-IDF를 이용해 추출한 질병 키워드들은 심방 세동, 심부전, 부정맥, 유방암, 류마티스 관절염, 복합 고지혈증, 다발성 경화증, 비알코올성 지방간염, 조영제 유발 신손상, 전립선암으로 이 키워드들의 아토바스타틴 적응증 여부를 확인하기 위하여 최신 연구초록을 고찰하였다.

심방 세동은 심방에서 발생되는 부정맥의 일종으로, 아토바스타틴은 심장 수술 전후나 전기 심장율동전환(Electrical cardioversion)시에 발생할 수 있는 심방 세동 발생 예방에 효과적이었다[15,16]. 심부전 환자에서도 아토바스타틴이 혈관내피전구세포의 활동을 개선시켜 혈관내피의 기능을 향상시켰으며, 고용량의 아토바스타틴은 심실 조기 수축이나 심실 빈맥과 같은 부정맥 합병증 발생 빈도를 줄여주었다[17,18].

류마티스 관절염 환자는 심혈관 질환 및 고지혈증 발생이 높으므로 예방을 위한 아토바스타틴 복용이 권고되었으며, 다발성 경화증 환자에서도 아토바스타틴의 항염증 작용으로 인한 보조치료제로써의 가능성에 대해 보고되었다[19,20]. 비알코올성 지방간염 환자에서도 아토바스타틴을 1년간 복용 시 지방간염 관련지표가 개선되었으며, 조영제 유발 신독성 발생 시 고용량의 아토바스타틴은 염증작용을 줄여 신장세포의 괴사를 줄여주었다[21,22].

또한 아토바스타틴은 암유전자 발현에 직접적인 영향을 주어 항암효과를 유발하는데 잠재적 유방암(Breast cancer)세포에서는 콜레스테롤 항상성(Homeostasis)을 개선시켜 유방암 유전자 발현을 억제하였고[23], 전립선암(Prostate cancer)세포에서는 아스피린(Aspirin) 혹은 당뇨병약인 메트포민(Metformin)과 병합요법 시 전립선 암세포의 성장이 억제된다고 보고되었다[24,25].

본 연구는 몇 가지 제한점을 가진다. 네트워크 분석은 확률할당에 의한 표본추출을 통하여 모집단을 추론하는 기존의 통계 분석법과는 달리 노드들의 관계에 중심을 둔 분석법으로 결과에 대한 일반화에 제한을 가진다. 이에 본 연구에서는 이를 극복하기 위하여 1990년부터 2017년 말까지 전수의 초록을 이용하여 분석하였다.

두 번째로 본 연구에서 추출된 키워드들을 임상적 평가하기 위하여 최신 연구 초록을 고찰하였는데 고찰에 제외된 초록들 중에 메타분석 및 체계적 문헌 고찰을 이용하여 해당 질병에 대한 적응증으로 결론을 내린 경우도 있었을 것으로 생각된다. 또한 출판된 연구결과만 이용함으로 인하여 출판되지 않은 연구결과들에 의해 추출된 키워드

들의 평가가 바뀔 오류도 가진다. 마지막으로 본 연구에서는 추출된 키워드 전처리 과정에서 질병 키워드 추출 시 인적 에러에 대한 영향도 고려해야 할 것이다.

결론적으로 텍스트 마이닝 및 네트워크 분석과 같은 비정형 데이터 분석을 통하여 대규모 의학 정보에서 기존 약물의 새로운 적응증인 질병들을 추출할 수 있었다. 이러한 비정형 데이터 분석은 향후 신약 재창출 산업의 전략 중 하나로 활용될 수 있을 것으로 기대한다.

ORCID

Hwee-Soo Jeong *https://orcid.org/0000-0001-5937-2827*

REFERENCES

1. Lee JY, Park BJ. A method to evaluate scientific evidence in clinical preventive medicine. J Korean Med Assoc 2011;54(10):1006-1012.
2. Park HA. Evidence base medicine and pre-appraised resources. Korean J Family Med 2010;31:897-903.
3. Song MH, Park DK, Lee YH. Medical informatics methods for the clinical evidence extraction. J Korean Med Assoc 2012;55(8):741-747.
4. The Korean Society of Medical Informatics Education. Healthcare informatics the 3rd. Edufactory: Seoul, 2015, p. 186-258.
5. Kim HH, Rhew KY. Analysis of adverse drug reaction reports using text mining. Korean J Clin Pharm 2017;27(4):221-227.
6. Greco I, Day N, Riddoch-Contreras J, Reed J, Soininen H, Kłoszewska I, et al. Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. J Transl Med 2012;10:217.
7. Meng G, Meng X, Ma X, Zhang G, Hu X, Jin A, et al. Application of ferulic acid for Alzheimer's disease: combination of text mining and experimental validation. Front Neuroinform 2018;12:31.
8. Salami JA, Warraich HJ, Valero-Elizondo J, Spatz ES, Desai NR, Rana JS, et al. National trends in nonstatin use and expenditures among the US adult population from 2002 to 2013: insights from medical expenditure panel survey. J Am Heart Assoc 2018;7(2):e007132.
9. Bae G, Park C, Lee H, Han E, Kim DS, Jang S. Effective policy initiatives to constrain lipid-lowering drug expenditure growth in South Korea. BMC Health Serv Res 2014;14:100.
10. Mach F, Ray KK, Wiklund O, Corsini A, Catapano AL, Bruckert E, et al. Adverse effects of statin therapy: perception vs. the evidence - focus on glucose homeostasis, cognitive, renal and hepatic function, haemorrhagic stroke and cataract. Eur Heart J 2018;39(27):2526-2539.
11. Lee SS. Network analysis methods applications and limitation. Crbooks: Seoul, 2018, p. 32-36.
12. Kwank KY. Social network analysis. Crbooks: Seoul, 2017, p. 2-230.
13. Song M. Text mining. Crbooks: Seoul, 2017, p. 105-121.
14. Pfizer Inc. LIPITOR-atorvastatin calcium trihydrate tablet, film coated. Available at http://labeling.pfizer.com/ShowLabeling.aspx?id=587#section-6 [Accessed on June 26, 2018].
15. Yuan X, Du J, Liu Q, Zhang L. Defining the role of perioperative statin treatment in patients after cardiac surgery: a meta-analysis and systematic review of 20 randomized controlled trials. Int J Cardiol 2017;228:958-966.
16. Yan P, Dong P, Li Z, Cheng J. Statin therapy decreased the recurrence frequency of atrial fibrillation after electrical cardioversion: a meta-analysis. Med Sci Monit 2014;20:2753-2758.
17. Oikonomou E, Siasos G, Zaromitidou M, Hatzis G, Mourouzis K, Chrysohoou C, et al. Atorvastatin treatment improves endothelial function through endothelial progenitor cells mobilization in ischemic heart failure patients. Atherosclerosis 2015;238(2):159-164.
18. He XZ, Zhou SH, Wan XH, Wang HY, Zhong QH, Xue JF. The effect of early and intensive statin therapy on ventricular premature beat or non-sustained ventricular tachycardia in patients with acute coronary syndrome. Cardiol J 2010;17(4):381-385.
19. Naerr GW, Rein P, Saely CH, Drexel H. Effects of synthetic and biological disease modifying antirheumatic drugs on lipid and lipoprotein parameters in patients with rheumatoid arthritis. Vascul Pharmacol 2016;81:22-30.
20. Li XL, Zhang ZC, Zhang B, Jiang H, Yu CM, Zhang WJ, et al. Atorvastatin calcium in combination with methylprednisolone for the treatment of multiple sclerosis relapse. Int Immunopharmacol 2014;23(2):546-549.
21. Takeuchi M, Sakasai-Sakai A, Takata T, Ueda T, Takino J, Tsutsumi M, et al. Serum levels of toxic AGEs (TAGE) may be a promising novel biomarker in development and progression of NASH. Med Hypotheses 2015;84(5):490-493.
22. He X, Li L, Tan H, Chen J, Zhou Y. Atorvastatin attenuates contrast-induced nephropathy by modulating inflammatory responses through the regulation of JNK/p38/Hsp27 expression. J Pharmacol Sci 2016;131(1):18-27.
23. Kimbung S, Chang CY, Bendahl PO, Dubois L, Thompson JW, Mc-

Donnell DP. et al. Impact of 27-hydroxylase (CYP27A1) and 27-hydroxycholesterol in breast cancer. Endocr Relat Cancer 2017;24(7): 339-349.

24. Wang ZS, Huang HR,, Zhang LY, Kim S, He Y, Li DL, et al. Mechanistic study of inhibitory effects of metformin and atorvastatin in combination on prostate cancer cells in vitro and in vivo. Biol Pharm Bull 2017;40(8):1247-1254.
25. He Y, Huang H, Farischon C, Li D, Du Z, Zhang K, et al. Combined effects of atorvastatin and aspirin on growth and apoptosis in human prostate cancer cells. Oncol Rep 2017;37(2):953-960.

KCI