대규모 언어 모델 (LLM) 기반 COVID-19 백신 접종자별 기저질환 및 알러지 텍스트 데이터 추출 방법에 대한 연구

김솔아[1], 윤지석[2], 정다은[1], 김호영[2], 조석헌*
[1]전북대학교, [2]국립금오공과대학교, *University of California, San Diego (UCSD)

codkan20@gmail.com, jiseok1234@gmail.com, ddueun0525@gmail.com,

knr3908@gmail.com, *justinshcho@gmail.com

Research on Text Data Extraction for Medical History and Allergy of COVID-19 Vaccine Recipients based on Large Language Model (LLM)

Sola Kim[1], Jiseok Yoon[2], Daeun Jeong[1], Hoyoung Kim[2], Seokheon Cho*
[1]Jeonbuk National University, [2]Kumoh National Institute of Technology,
*University of California, San Diego (UCSD)

요 약

There has been a significant increase in reported cases of adverse vaccine effects due to the unprecedentedly high vaccination rates resulting from the COVID-19 pandemic. To efficiently manage this vast amount of information concerning not only COVID-19 vaccines but also various other vaccinations, the United States operates and manages the Vaccine Adverse Event Reporting System (VAERS). In this study, we proposed methods for extracting and preprocessing text data on medical history and allergies from the VAERS dataset, which can be used to predict post-vaccination adverse effects of COVID-19 vaccines. We structured the text data, grouped individual medical history as well as allergy, and thus created a dataset reflecting individual characteristics by utilizing both large language model (LLM) and various text extraction algorithms. Extracted text data on medical history and allergies can facilitate the understanding of COVID-19 adverse effect of vaccines and serve as key data for effectively responding future adverse reactions.

Ⅰ. 서 론

COVID-19 팬데믹 이후, 전 세계적으로 집단 면역 (herd immunity) 형성과 개인의 생명을 보호하기 위해 이례적으로 높은 백신 접종률을 기록했다 [1]. 하지만, COVID-19 백신 접종 결과 보고에 따르면 백신 접종 후 사망을 포함한 다양한 부작용 피해가 발생함을 알 수 있다 [2]. 이러한 COVID-19 백신의 부작용에 대한 중요성을 인식하고 이해를 높이기 위해, 각 개인의 기저질환 및 알러지와 백신 접종 후 발생할 수 있는 다양한 부작용 증상들 사이의 연관성을 파악하는 것이 필요하다. 특히, 미국에서는 Vaccine Adverse Event Reporting System (VAERS)을 통해 백신 접종자들의 기본 정보, 기저질환, 알러지 그리고 부작용 증상들과 같은 다양한 데이터들을 수집 및 관리를 하고 있다 [3]. 그러나 VAERS 에서 제공하는 데이터셋에는 복잡한 텍스트형 데이터들이 포함되어 있기 때문에, COVID-19 백신 접종자들의 기저질환 및 알러지와 부작용 증상 간의 연관성을 분석하기 전에, 먼저 해당 데이터셋에서 텍스트 정보를 추출하는 작업이 선행되어야 한다.

따라서 본 연구에서는 VAERS 가 제공하는 COVID-19 백신 접종자들의 기저질환과 알러지 데이터를 대규모 언어 모델 (Large Language Model: LLM)을 이용하여 효율적으로 텍스트들을 추출하고 리스트화 하고자 한다.

Martuza Ahamad *et al.*은 Vaccine Adverse Event Reporting System (VAERS) 데이터를 사용하여 COVID-19 백신 접종 이후 발생할 수 있는 상황에 대해 주목하였다 [4]. 문자열 매칭 (string matching) 및 키워드 선택 (keyword selection) 기법을 적용하여 환자 개인의 기저질환과 백신 접종 후 부작용 증상을 추출하고 그를 바탕으로 COVID-19 백신 접종 이후 발생할 수 있는 환자의 사망 여부, COVID-19 양성 판정 여부 및 입원 여부 상태를 예측하는 모델을 개발하였다. Bosung Kim *et al.*은 백신 안전 모니터링 시스템의 개발과 성능 향상을 목표로 백신 부작용 증상에 대한 텍스트 추출 연구를 수행하였다 [5, 6]. 국제의약용어 (Medical Dictionary for Regulatory Activities: Med DRA)를 [7] 기반으로 VAERS 데이터에 GPT (Generative Pre-trained Transformer)-3 와 같은 자연어 처리 모델 (Natural Language Processing Model: NLP)을 사용하여 백신 부작용 증상들과 각 증상에 대한 백그라운드 지식을 포함하는 SYMPTOMIFY 데이터셋을 소개하였다. Saeyeon Cheon *et al.*은 VAERS 데이터셋에서 COVID-19 백신 접종자들의 기저질환과 알러지 정보에 대하여 NLP 중의 하나인 Word2Vec 을 사용하여 단어를 벡터화하고 차원을 줄였다. 또한 비지도 기계 학습 (unsupervised machine learning)인 DBSCAN 알고리즘을 활용하여 증상을

클러스터링하고 각 증상 클러스터의 특성을 분석했다. 부작용 간의 연관 규칙을 발견하기 위하여 Apriori 알고리즘을 통한 데이터 마이닝 접근 방식을 사용하였다 [8].

선행 연구들에서는 COVID-19 백신 접종자들의 부작용에 대한 텍스트 추출을 주로 대상으로 하거나 사전 학습 언어 모델 (Pre-trained Language Model)이 아닌 NLP 를 사용하여 기저 질환과 알러지에 대한 텍스트 추출 방법을 제시하였다. 하지만, 본 연구에서는 텍스트 기반 데이터에서 주로 기저 질환과 알러지를 추출하는 데 중점을 두었다. 이를 위해 대규모 언어 모델 (Large Language Model)과 다양한 추출 알고리즘을 활용한다. 추출된 기저 질환과 알러지 데이터는 향후 COVID-19 부작용을 예측하는 모델을 구축하는데 중요한 정보로 활용될 수 있다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 본 연구에 사용된 VAERS 데이터셋에 대해 설명한다. 제 3 장에서는 이 데이터셋에 포함된 COVID-19 백신 접종자들의 기저질환과 알러지 정보를 추출하는 대규모 언어 모델 기반의 텍스트 추출 방법을 제시한다. 마지막으로, 제 4 장에서는 결론과 향후 연구 방향에 대해 논의한다.

Ⅱ. Vaccine Adverse Event Reporting System 데이터셋

본 연구는 Vaccine Adverse Event Reporting System (VAERS)에서 제공하는 2021 년 데이터를 이용하여 COVID-19 백신 접종자들이 가지고 있는 기저질환과 알러지 정보들에 대해서 텍스트 데이터 추출을 수행하였다 [3]. VAERS 는 백신 부작용과 관련된 데이터들을 수집하여 제공하는 미국의 시스템이다. VAERS 에서 제공되는 2021 년 데이터셋은 크게 2021VAERSDATA, 2021VAERSVAX 그리고 2021VAERSSYMPTOMS 3 개로 구성되어 있다. 이들은 모두 2021 년 1 월 1 일부터 2021 년 12 월 31 일까지의 기간에 수집한 데이터를 포함하고 있다.

Bosung Kim *et al.*은 2021VAERSDATA 데이터셋에 자연어 처리 모델 (Natural Language Processing Model) 중의 하나인 GPT (Generative Pre-trained Transformer)-3 를 적용하여 백신 접종자들에게 발생한 다양한 부작용 증상들의 텍스트를 추출하였다 [5]. 그러나 백신 접종자들이 이미 가지고 있는 만성 또는 장기적인 건강 상태를 나타내는 기저질환 (Medical History)들과 복용한 약물이나 음식 또는 기타 제품 등에 대한 알러지들 (Allergies) 정보에 대하여 사전 학습 언어 모델 (Pre-trained Language Model)을 사용하여 텍스트 추출을 수행한 선행 연구들은 충분하지 않다. 또한, 백신 접종자마다 매핑되어 저장된 기저 질환과 알러지 정보 각각이 최대 32,000 바이트의 문자형으로 되어 있어 전처리가 복잡하다는 문제점이 있다.

Ⅲ. 대규모 언어 모델을 이용한 텍스트 데이터 추출

대규모 언어 모델 (Large Language Model: LLM) 은 대규모의 텍스트 데이터를 사용하여 학습된 인공지능 모델이다. 이는 자연어 처리와 관련된 작업에 사용되는데, 텍스트의 의미와 문맥을 이해하고, 문장 생성, 번역, 요약 및 질문 응답 등 다양한 자연어 이해 및 생성 작업을 수행할 수 있다.

그림 1 은 2021VAERSDATA 에 포함되어 있는 기저질환과 알러지에 대하여 텍스트 추출하는 과정을 보여준다. 우선, 최대 32,000 바이트의 문자형으로 되어있는 기저질환 (Medical History)과 알러지 (Allergies) 데이터를 소문자로 변환한다. 그리고 연결어 와 같은 표현을 ‘,’ 형태로 변경하고 ‘,’에 의하여 구분된 단위로 데이터를 토큰화 (Tokenization)했다. 이후 정규표현식 ‘^a-zA-Z0-9-’ 을 사용하여 불필요한 특수 문자를 제거하였으며, ‘na, ’‘nan, ’‘none, ’‘none we are aware of, ’‘n/a, ’ ‘no, ’‘none known, ’‘no reported, ’ ‘NKDA’ 등과 같이 결측치를 표현하는 단어 목록을 집합화하여 모두 ‘null’ 값으로 변환했다. 기저질환과 알러지에 대한 텍스트를 추출하는데 있어서 여기까지의 과정들은 동일하게 적용하였다. 하지만, 기저질환과 알러지 데이터에 적용할 이후의 과정들은 상이하다.

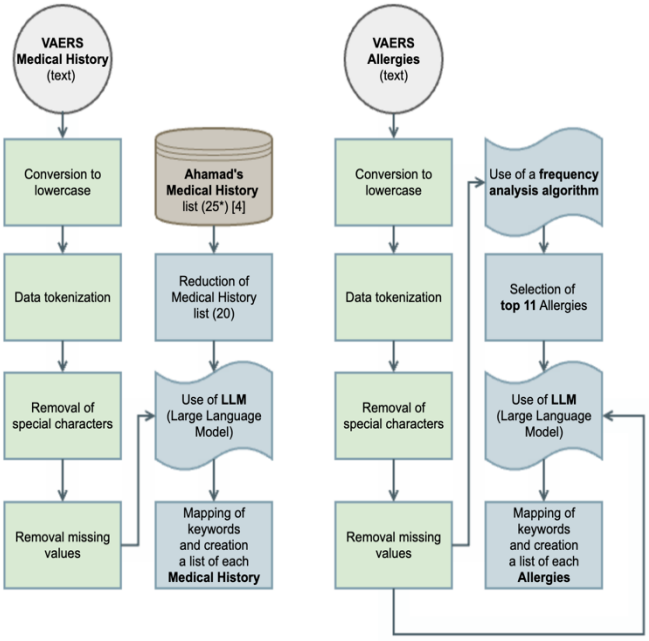

그림 1. 대규모 언어 모델 기반 기저질환 및 알러지 데이터 추출 방법 절차

결측값 처리 이후 기저질환 정보에 대한 텍스트 추출 과정은 다음과 같다. Martuza Ahamad *et al.*은 의학 전문가들로부터 자문을 통해 COVID-19 과 관련있는 기저질환 25 개를 선정하였다 [4]. 선정된 기저질환은 총 25 개였지만, 유사한 항목들을 통합하는 과정을 거쳐 20 개의 기저질환으로 축소하였다. 축소한 기저질환들은 다음과 같다: 다른 약 복용 (Taking other medicine)과 이전 백신 접종 기록 (Prior vaccine)은 각각 하나의 범주로서 처리하기에 수많은 세부 사항들을 고려해야 하므로 제거하였다. 알러지 과거 정보 (Allergic history)는 2021VAERSDATA 에 포함된 알러지 데이터와 중복되기 때문에 고려하지 않았다. 또한, 고혈압 (hypertension)은 비정상 혈압 (Abnormal blood pressure)에 포함되기 때문에 비정상 혈압이라는 기저질환 키워드에 종속시켰다. 고지질혈증 (Hyperlipidemia)과 고 콜레스테롤 (High cholesterol)은 비슷한 의미이기 때문에 고지질혈증을 고 콜레스테롤 이라는 기저질환 키워드에 포함시켰다.

표 1 은 이러한 20 개의 기저질환들의 리스트를 보여준다. 또한, 각 기저질환 별 데이터 샘플 수뿐만 아니라 개별 기저질환 에 매핑되어 포함되는 대표 키워드들을 나타낸다. 예를 들면, 당뇨병 (Diabetes)을 기저질환으로 가지고 있는 총 데이터 샘플 개수는 25,937 개이다. 또한, ‘type 2 diabetes,’ ‘dm,’ ‘t2dm’ 등의 키워드들은 당뇨병을 다양하게 표현하는 키워드들이기 때문에, 해당 키워드들 은 당뇨병 (Diabetes)으로 종속하였다.

이후 최종적으로 선정된 20 개의 기저질환을 대규모 언어 모델 (LLM)인 GPT-4 를 [8] 활용하여 토큰화된 데이터와의 키워드 매핑을 진행했다. GPT-4 에서 사용된 Prompt 는 다음과 같다:

"Extract all words from the {data tokenized from the list of VAERS Medical History}, which indicate the same disease as {Ahamad's Medical History}."

표 1. GPT-4 를 이용한 기저질환 리스트 및 개별 기저질환에 대한 키워드 매핑

Medical History (Sample count)	List of Keywords
Diabetes (25,937)	{'diabetes', 'type 2 diabetes', 'dm', 't2dm', …}
Abnormal Blood Pressure (54,388)	{'hypotension', 'high blood pressure', 'hypertension', …}
High Cholesterol (21,672)	['hyperlipidemia', 'dyslipidemia', …]
Arthritis (19,221)	['arthritis', 'chronic arthritis', 'jia', …]
Asthma (28,463)	['asthma', 'asthma multiple sclerosis', …]
Migraine (10,616)	['migraines', 'migraine', …]
Copd (6,627)	['copd', 'chronic bronchitis', …]
Gerd (12,233)	['gerd', acid reflux', 'nerd', …]
Anxiety (15,134)	['anxiety', 'anxiety disorder', 'ptsd', …]
Obesity (10,708)	['obesity', 'overweight', 'calories', …]
Depression (15,287)	['depression', 'depressive disorder', …]
Thyroid Disorder (24,832)	['hypothyroidism', 'hypothyroid', …]
Anemia (4,595)	['anemia', 'iron deficiency anemia', …]
Dementia (2,164)	['dementia', 'alzheimer's', …]
Cancer (11,784)	['cancer', 'lung cancer', 'myeloma', …]
Kidney Disease (7,130)	['renal', 'hypertensive nephropathy', …]
Heart Disease (53,022)	['mi', 'aortic stenosis', 'rheumatic' …]
COVID-19 Positive History (7,678)	['covid-19 positive history', …]
Atrial Fibrillation (6,143)	['atrial fibrillation', 'paroxysmal', …]
Pain Symptoms (28,083)	['pain', 'chronic pain', 'back pain', …]

그림 2 는 표 1 에서 열거한 20 개의 기저질환들마다 매핑되는 키워드들의 수를 보여준다. 예를 들면, 당뇨병과 고 콜레스테롤에 매핑되는 키워드들의 개수는 각각 60 개와 45 개이다. 총 680 개의 토큰화된 키워드들이 20 개의 기저질환들에 매핑된다.

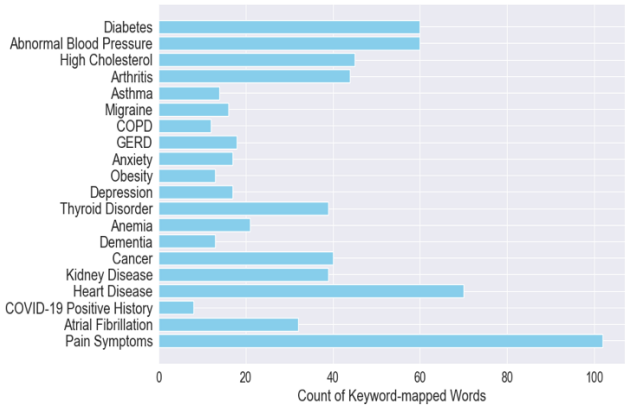

그림 2. 개별 기저질환에 대한 매핑되는 키워드 개수

결측값 처리 이후 알러지 정보에 대한 텍스트 추출 과정은 다음과 같다. 토큰화된 데이터들에 대하여 빈도 분석 알고리즘 (Frequency Analysis Algorithm)을 사용하여 빈도수가 높은 순서대로 11 개의 알러지들을 선정하였다. 표 2 는 이러한 11 개의 알러지들의 리스트를 보여준다. 또한, 각 알러지 별 데이터 샘플 수 그리고 개별 알러지에 매핑되어 포함되는 대표 키워드들을 나타낸다. 이후 선정된 11 개의 알러지들을 GPT-4 를 활용하여 토큰화된 데이터와의 키워드 매핑을 진행했다. GPT-4 에서 사용된 Prompt 는 다음과 같다:

"Extract all words from the {data tokenized from the list of VAERS Allergies}, which fall under the category of {top 11 Allergies}."

표 2. GPT-4 를 이용한 알러지 리스트 및 개별 알러지에 대한 키워드 매핑

Allergies (Sample count)	List of Keywords
Sulfa (28,903)	['sulfa', 'sulfa drugs', 'sulfonamide', …]
Penicillin (34,467)	['penicillin', 'pcn', 'penicillins', …]
Food_allergies (18,245)	['banana', 'peanut', 'wheat', …]
Seafood_and_Shellfish (9,346)	['shrimp', 'lobster', 'crab', 'squid', …]
Morphine (5,238)	{'morphine', 'morphia', 'morphinum', …}
Codeine (9,968)	{'codeline', 'codelin', 'codeine', …}
Latex & Rubber (9,762)	['latex', 'rubber']
Amoxicillin (7,451)	['amoxicillin', 'amoxi']
Erythromycin (3,876)	['erythromycin']
Dust (3,091)	['dust']
Mold (2,568)	['mold', 'mite', 'dustmite']

그림 3 은 표 2 에서 열거한 11 개의 알러지들마다 매핑되는 키워드들의 수를 보여준다. 예를 들면, 항생제 (Sulfa)와 페니실린 (Penicillin)에 매핑되는 키워드들의 개수는 각각 16 개와 12 개이다. 총 144 개의 토큰화된 키워드들이 11 개의 알러지들에 매핑된다.

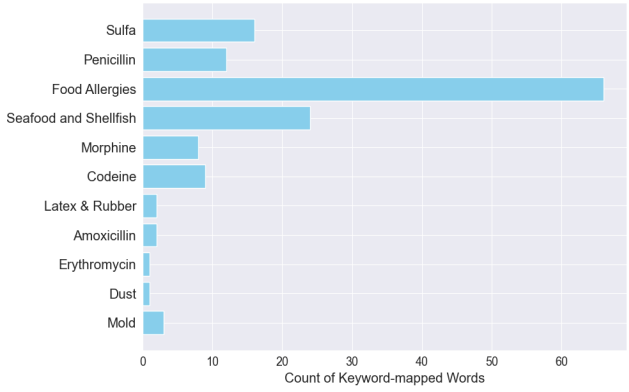

그림 3. 개별 알러지에 대한 매핑되는 키워드 개수

그림 1 에서 도시하는 것처럼 2021VAERSDATA 에 포함된 COVID-19 백신 접종자들이 가지고 있는 기저질환과 알러지 데이터들에 대한 텍스트 추출을 위해 대규모 언어 모델을 적용하면 최종적으로 537,218 개의 데이터 샘플을 효율적으로 얻을 수 있다. 이렇게 GPT-4 를 통해 텍스트 추출 결과물들을 다음 링크를 통해서 얻을 수 있다: (https://github.com/ffe4el/KICS_LLM/tree/main)

- 20 개의 기저질환 리스트 및 해당 키워드
- 11 개의 알러지 리스트 및 해당 키워드
- COVID-19 백신 접종자들의 데이터

IV. 결론

본 연구에서는 COVID-19 백신 접종자들의 기저질환 및 알러지 데이터를 추출하였다. 텍스트형 데이터인 기저질환과 알러지 데이터를 추출하기 위해서 대규모 언어모델 (Large Language Model: LLM)과 다양한 추출 알고리즘을 사용하여 추후 인공지능 모델이 학습할 수 있는 형태로 가공하였다. 텍스트형 데이터를 토큰화하여 LLM 중 하나인 GPT-4 를 이용해 적절한 범주의 기저질환 및 알러지로 그룹화하였다.

이후에는 COVID-19 백신 접종 후 부작용에 대한 효율적인 대처를 미리 제시하고자 본 연구를 통해 추출한 기저질환과 알러지를 특성으로 이용하여 개개인의 부작용 증상 발생 가능성을 예측하는 모델로 확대하는 향후 연구가 필요하다.

ACKNOWLEDGMENT

Following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation) in 2024 (2022-0-01067)

참 고 문 헌

[1] Freda Kreier, “Ten Billion COVID Vaccinations: World Hits New Milestone,” Nature, Jan. 2022.

[2] Bruna A. S. Machado, Katharine V. S. Hodel, Larissa M. D. S. Fonseca, Vinícius C. Pires, Luis A. B. Mascarenhas, Leone P. C. da S. Andrade, Marcelo A. Moret, and Roberto Badaró, “The Importance of Vaccination in the Context of the COVID-19 Pandemic: A Brief Update Regarding the Use of Vaccines,” Vaccines (Basel), vol. 10, no. 4, Apr. 2022.

[3] VAERS: Vaccine Adverse Event Reporting System, 2021., Available: https://vaers.hhs.gov/data/datasets.html

[4] Martuza Ahamad *et al.*, “Adverse Effects of COVID-19 Vaccination: Machine Learning and Statistical Approach to Identify and Classify Incidences of Morbidity and Postvaccination Reactogenicity,” Healthcare (Basel), vol. 11, no. 1, Dec. 2022.

[5] Bosung Kim and Ndapa Nakashole, “SYMPTOMIFY: Transforming Symptom Annotations with Language Model Knowledge Harvesting,” Association for Computational Linguistics, pp. 11667- 11681, Dec. 2023.

[6] Bosung Kim and Ndapa Nakashole, “Data Augmentation for Rare Symptoms in Vaccine Side-Effect Detection,” Association for Computational Linguistics., pp. 310- 315, May. 2022.

[7] MedDRA: Medical Dictionary for Regulatory Activities Available: https://www.ich.org/page/meddra

[8] Saeyeon Cheon, Thanin Methiyothin, and Insung Ahn, “Analysis of COVID-19 Vaccine Adverse Event Using Language Model and Unsupervised Machine Learning,” PLoS ONE, vol. 18, no. 2, Feb. 2023.

[9] OpenAI, “GPT-4 Technical Report,” arXiv, Mar. 2023.