ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
*J Korean Soc Qual Manag* Vol. 50, No.4: 717-734, December 2022
*https://dx.doi.org/10.7469/JKSQM.2022.50.4.717*

# 우수 의약품 제조 기준 위반 패턴 인식을 위한 연관규칙과 텍스트 마이닝 기반 t-SNE분석

이준오* · 손소영**†

* 연세대학교 산업공학과 석사과정생
** 연세대학교 산업공학과 교수

# Violation Pattern Analysis for Good Manufacturing Practice for Medicine using t-SNE Based on Association Rule and Text Mining

Jun-O Lee* · So Young Sohn**†

* Master Course, Dept. of Industrial Engineering, Yonsei University
** Professor, Dept. of Industrial Engineering, Yonsei University

## ABSTRACT

**Purpose:** The purpose of this study is to effectively detect violations that occur simultaneously against Good Manufacturing Practice, which were concealed by drug manufacturers.

**Methods:** In this study, we present an analysis framework for analyzing regulatory violation patterns using Association Rule Mining (ARM), Text Mining, and t-distributed Stochastic Neighbor Embedding (t-SNE) to increase the effectiveness of on-site inspection.

**Results:** A number of simultaneous violation patterns was discovered by applying Association Rule Mining to FDA's inspection data collected from October 2008 to February 2022. Among them there were 'concurrent violation patterns' derived from similar regulatory ranges of two or more regulations. These patterns do not help to predict violations that simultaneously appear but belong to different regulations. Those unnecessary patterns were excluded by applying t-SNE based on text-mining.

**Conclusion:** Our proposed approach enables the recognition of simultaneous violation patterns during the on-site inspection. It is expected to decrease the detection time by increasing the likelihood of finding intentionally concealed violations.

**Key Words:** Regulatory Inspection, Good Manufacturing Practice, Text Mining, Association Rule Mining, t-SNE(t-distributed Stochastic Neighbor Embedding)

● Received 10 September 2022, 1st revised 5 October 2022, accepted 1 November 2022
† Corresponding Author(sohns@yonsei.ac.kr)

* 교신저자의 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2020R1A2C2005026)

# 1. 서 론

약을 처방받거나 처방하는 사람들은 의약품의 제조 단계에서 기인한 결함 여부를 감지할 가능성이 매우 희박하다. 의약품의 품질은 약물 치료 환자의 안전과 직접적으로 연관되어 있기 때문에 규제 당국에 의하여 엄격히 관리되어야 할 필요성이 있다. 의약품 품질 시스템은 국제 표준 기구(International Standards Organization, ISO)의 품질 개념을 기반으로 구성되어 있으며, 적정한 의약품 품질 시스템의 효과성은 제조 현장에서의 실태조사를 통해 평가될 수 있다(ICH, 2008; PIC/S, 2009). 의약품 제조소에 대한 실태조사는 문제를 발굴하고 개선을 이끌어내기 위한 필수적인 프로세스로서 의약품 제조업체에 대한 사회적 신뢰를 형성하기 위해 규제 당국에 의해 수행된다(Kumar et al., 2013; Choi & Park, 2018). 우리나라는 2014년 7월 국제 기구인 의약품 품질 상호 협력 기구(PIC/S : Pharmaceutical Inspection Co-Operation Scheme)에 가입하고 해당 기구의 우수 의약품 제조 기준(GMP: Good Manufacturing Practice)에 근거하여 의약품 제조소를 대상으로 품질경영, 시설장비, 제조, 시험실, 원자재, 포장표시의 6개 GMP 감시 분야에 대해 실태조사를 수행하고 있다. 실태조사는 크게 사전 실사(Pre Approval inspection)와 사후 실사(Post Approval Inspection)의 두 가지 방식으로 이루어진다. 사전 실사는 의약품 생산을 위한 제조소 승인을 위해 진행되는 실태 조사이다. 사후 실사는 GMP 적합판정서의 갱신을 위해 주기적으로 이루어지는 정기 실태조사와 시정사항 조치 결과를 확인하기 위해 이루어지는 재실사 그리고 신고 등의 특별한 경우가 발생했을 때 원인을 확인하기 위해 실시하는 특별점검으로 이루어져 있다. 실태조사 과정에서 규정 위반사항이 있을 경우 규제 당국은 제조소에 대해 보완 지시, 시정 이행 결과 확인의 과정을 거쳐 GMP 기준 적합 여부를 결정하게 된다. 시정 및 보완 조치와는 별개로 위반사항이 행정 처분의 기준에 해당할 경우 제조업무정지 등의 처분이 내려지게 된다. 이러한 실태조사가 수행됨에도 불구하고 최근 몇 년 동안 의약품 결함에 의한 회수 조치는 증가하는 경향을 보이고 있다. 이는 우수 의약품 제조 기준(GMP) 위반에 의한 제품 결함 발생을 감소시키기 위해 수행되는 실태조사의 기존 방법에 대해 재고할 필요가 있다는 것을 시사한다.

실태조사를 정기적으로 수행하는 방법의 한계점은 실태조사가 상호 약속된 일정에 수행되기 때문에 수검을 받는 제조소가 실태조사 일정에 앞서 의도적으로 위반 사항을 은폐할 경우 규정 위반 실태를 적발하기 어렵다는 것이다. 현재 우수 의약품 제조 기준(GMP)을 만족시키기 위해 회사 차원에서 많은 비용이 드는 법, 규정 전문가와 기술 자문 전문가를 두는 것이 권장되기도 하며 최소한의 시설, 장비, 품질 시스템을 구축하고 운영하기 위해서는 지속적인 투자가 필요하다. 따라서 충분한 자원을 갖지 못한 제약회사에서 요구사항을 충족시키는 것은 매우 어려운 문제이다(Abdellah et al., 2016; Lee et al., 2022). 기준을 만족시키기 위해서는 회사의 재정 규모에 비해 과다한 자원 투입이 필요할 수 있으며 이에, 자원 투입 효율화에 힘쓰기보다는 왜곡된 방법을 선택하여 해결하는 경우가 있다. 제조 시스템을 정비하지 않고 규제 당국의 정기 실태조사 시 미흡한 부분을 은폐하여 실태조사 적합 판정을 받는 경우이다. 실제 2021년 식품의약품안전처를 대상으로 진행된 국정감사에서 규제당국과 의약품 제조소간 실태조사 일정을 사전에 조율하고 진행하는 '정기 실태조사'의 2018~2020년 위반 적발율은 불시에 제조소를 점검하는 '특별점검'의 위반 적발율보다 현저히 낮은 것으로 확인됐다. 이에 불시에 이루어지는 특별점검의 확대가 주장되기도 하지만 불시에 이루어지는 실태조사의 무분별한 확대는 기업의 권리를 침해할 소지가 있다. 실태조사에 제한된 인원과 시간이 투입되기 때문에 대규모 제조공장 전체에 대한 심층 점검이 어렵다는 점도 실태조사의 한계점이다. 보통 2명이 3일에서 5일 정도로 제조소 점검을 수행하고 있다(PIC/S, 2009). 제한된 시간 내에 많은 현장 조사가 요구된다. 투입되는 공수가 유한한 한계를 극복하고자 GMP 감시가 필요한 분야 중에서 실태조사 필요도가 높은 점검 분야를 사전에

선정하고, 우선순위를 부여하여 집중된 실태조사가 이루어지게 하고자 위험성 평가 도구(Risk Assessment Tool)가 사용되기도 한다. 제조 공정의 복잡성과 생산하고 있는 제품의 민감성, 제품이 갖는 사회적 중요성 그리고 해당 제조소의 과거 실태조사 관찰사항 정보 등이 위험성 평가에 활용된다(PIC/S, 2012). 제조소에서 규정 위반 가능성이 높은 특정 분야에 대해 점검 여력을 집중함으로써 제한된 공수를 효과적으로 활용할 수 있도록 한다. 이는 특정 분야의 실정을 면밀하게 관찰하는데 도움이 될 수는 있지만, 선정되지 않은 분야에서의 실태 조사는 소홀하게 이루어질 수 있는 위험이 있다. 또한 실태조사 중에 적발된 위반 사례를 토대로 감추어진 위반 실태를 효과적으로 적발하거나, 적발하는데 소요되는 시간을 단축하는 것에 대한 기여는 제한적이다. 동시에 발생하는 규정 위반 패턴에 대해 탐구는 적발된 위반 사례를 근거로 은닉된 위반 실태를 적발하는데 도움이 될 수 있다. 또한 패턴이 미리 제시되어 있다면 적발 소요 시간 단축에도 기여할 수 있다.

의약품 제조소에 대한 실태조사가 효과적으로 이루어지는데 기여하고자 여러 연구가 진행되었다. Lim & Cho.(2021)은 EU 국가별 GMP 부적합 판정율에 대한 영향 인자를 분석하기 위하여 EU GMP 실태조사 규정 위반 자료를 다중회귀분석에 적용하였다. Ban et al.(2019)은 미국과 유럽의 의약품 제조소를 중심으로 일정 기간 동안의 규정 별 위반 경향을 분석하여 어떠한 규정이 지속적으로 문제가 되는지 파악하고자 하였다. Linna et al.(2007)은 실태조사를 수행하기 전에 제조소에 대한 사전 정보를 최대한 많이 얻고자 델파이 방법을 활용한 설문 방식의 제조소 사전 점검 Tool을 개발하였다. 지금까지의 연구는 동시에 발생하는 규정 위반 패턴에 대한 탐구 보다는 개별 규정에서의 위반 경향과 GMP 최종 적합 판정에 관련되는 요인을 파악하는 것에 집중되어 있다.

본 연구에서는 연관규칙 마이닝과 텍스트 마이닝에 기반한 t-SNE(t-distributed Stochastic Neighbor Embedding) 분석을 통하여 동시에 발생하는 규정 위반 패턴을 효과적으로 발견하는 방법을 제시한다. 동시에 관찰되는 규정위반 패턴에는 규정이 정하고 있는 범위가 다른 규정과 유사하여 파생되는 동시 위반 패턴이 존재한다. 규정은 고립되어 존재하는 것이 아니라 내용이 상호 간 유기적으로 결부되어 구성되어 있기 때문에, 하나의 규정이 규율 하려는 사항이 다른 규정에서 제약하고 있는 내용과 중복되거나 연관되어 있는 경우가 많기 때문에 나타나는 패턴이다. 이것은 한 가지의 위반 행위에 대해 두 가지 이상의 위반 규정이 있는 것으로 동시적으로 나타나는 위반 행위를 예측하는데 도움이 되지 않는다. 이러한 패턴을 제외하는 것이 필요하다. 본 연구에서는 연관 규칙 마이닝을 활용하여 발굴된 패턴에서 규정 간의 상호 연관성에서 유래한 패턴을 제외하기 위해 텍스트 마이닝을 기반으로 한 t-SNE을 활용한다. 분석을 위한 데이터는 FDA의 규제업무사업소(ORA, Office of Regulatory Affairs)에서 제공하는 2008년 10월부터 2022년 02월 기간 동안의 실태조사 관찰사항(Inspection Observations Data)을 활용한다. 또한 규정의 내용을 확인하기 위해서 'FDA Medical Device Databases'에서 제공하는 CFR 규정집을 사용하였다.

본 연구의 규정 위반 패턴 인식 방법의 활용은 실태조사 중 규정 위반이 관찰됐을 경우 동시에 위반할 가능성이 높은 규정을 제시함으로써 제조소의 실정을 보다 더 잘 살펴볼 수 있게 하고 감추어진 사항에 대한 적발 가능성을 높이며 적발 소요 시간을 단축함으로써 실태조사의 목적이 잘 이행되도록 하는데 기여가 기대된다.

본 연구의 구성은 다음과 같다. 제2장에서 미국식품의약국(FDA)의 실태조사 과정과 본 연구에 사용된 방법론 분야에 대해서 고찰하였다. 이어서 3장에서는 기준 위반 패턴 분석을 위한 방법론을 제시하고 이어서 실증 분석하였다. 제4장에서는 본 연구의 결론과 시사점에 대해 논하였다.

# 2. 문헌 고찰

## 2.1. FDA의 의약품 제조소 실태조사

FDA는 미국 식품 의약국이다. FDA의 하위 조직인 규제업무사업소(ORA, Office Regulatory Affairs)에서 직접적으로 의약품 제조소에 대한 실태조사를 총괄하고 있다. FDA는 실태조사 중 규정 위반 사항이 관찰되면 제조소에 보완 요구 문서(Form 483)를 발행한다. 실태 조사 시 미국 연방 규정집(CFR : Code of Federal Regulation)에서 정하고 있는 우수 의약품 제조소(GMP) 규정이 위반 판단의 척도로 사용된다. 미국 연방 규정집(CFR)은 미국 연방행정부가 발행한 행정명령이다. 규제의 큰 범주를 나타내는 '타이틀(Title)'은 총 50개로 이루어져 있다. 각 타이틀안의 파트(Part) 번호는 특별한 규제영역을 다루고 있다. 21번의 타이틀에서는 식품과 의약품(Food and Drugs)과 관련된 규정을 다룬다. 21번 타이틀 안의 파트 211에는 완제의약품의 우수 의약품 제조소(GMP)에 대한 규제 항목이 실려 있다. Table 1에 나타난 바와 같이 21번 타이틀의 파트 211은 다시 11개의 서브 파트로 나뉘어 각각 일반 조항, 조직 및 인원, 건물 및 시설과 같은 특정 분야에 대해 규율한다. 또한 서브 파트를 구성하는 293개의 하위 규정은 규제하고 있는 내용을 상세하게 명시하고 있다.

**Table 1.** 21 CFR 211 Regulatory content by subpart

| | |
|---|---|
| Title 21<br>CFR 211 Part<br>Subpart | · Subpart A(CFR 211.1~211.3) : General Provisions<br>· Subpart B(CFR 211.22~211.34) : Organization and Personnel<br>· Subpart C(CFR 211.42~211.58) : Buildings and Facilities<br>· Subpart D(CFR 211.63~211.72) : Equipment<br>· Subpart E(CFR 211.80~211.94) : Control of Components and Drug Product Containers and Closures<br>· Subpart F(CFR 211.100~211.115) : Production and Process Controls<br>· Subpart G(CFR 211.122~211.137) : Packaging and Labeling Control<br>· Subpart H(CFR 211.142~211.150) : Holding and Distribution<br>· Subpart I(CFR 211.160~211.176) : Laboratory Controls<br>· Subpart J(CFR 211.180~211.198) : Records and Reports<br>· Subpart K(CFR 211.204~211.208) : Returned and Salvaged Drug Products |

## 2.2. 방법론 관련연구

대용량의 데이터를 대상으로 하는 연관규칙 분석의 결과에서 나타나는 많은 패턴으로부터 사용자에게 효용성이 높은 규칙을 판별하기 위해서는 명확한 해석과 판단을 필요로 한다(Kim et al., 2021). 효용성이 높은 규칙을 도출하는 방법에 대한 위한 연구로 Wang et al. (2004)은 아이템에 거래량을 가중치로 할당하는 방법으로 일정량 이상의 A제품을 구매한 고객이 B제품을 구매한다는 식의 보다 구체적인 정보를 얻을 수 있는 Weighted Association Rule Mining을 제안하였다. 아이템에 직접 Weight를 부여할 경우 발생할 수 있는 Apriori-Algorithm의 하향폐쇄성질(Downward closure property)의 위반을 해결하기 위하여 Tao et al.(2003)은 트랜잭션에 가중치를 부여하는 방법을 제시하였다. Han, Fu(1995)은 이미 계층 구조를 갖는 대규모의 데이터에서 하향식 점진적 연관성 분석을 통해 구체적인 규칙 정보를 발견할 수 있는 방법에 대해 논의하였다. Lin et al.(2002)은 대용량의 데이터에서 패턴이 폭

발적으로 도출되는 것을 방지하기 위하여 총 규칙의 개수를 제한하는 방식의 ASARM(Adaptive-Support Association Rule Mining)을 제안하였다.

TF-IDF(Term Frequency - Inverse Document Frequency)를 통해 단어 기준으로 표현된 문서의 유사도를 활용하여 주제가 비슷한 문서 별로 분류하려는 시도는 지속적으로 이루어져 왔다. Seong et al.(2022)은 외부 회계감사 자료에 대해 TF-IDF를 기반으로 한 문서 분류 방법을 사용한다. Bafna et al.(2018)은 TF-IDF를 사용하여 문서를 벡터화하고, K-Means와 Hierarchical clustering를 통해 군집화 하여 분류하는 방법을 제안한다. 분류 전 군집의 수와 주제를 미리 설정하기 위해서 LDA(Latent Dirchlet Allocation)가 사용되기도 한다(Onan et al., 2017).

t-SNE는 고차원의 데이터를 2차원 또는 3차원으로 축소하여 시각화 하는데 유용하여 다양한 분야에서 사용되고 있다. 최근 유전자 정보에 대한 높은 관심 속에 실험 기술의 급속한 성장으로 유전자 데이터를 얻는 것을 가능하게 했다. 유전자가 갖는 방대한 데이터를 시각화 하여 분석하기 위해 t-SNE이 주로 사용된다(Han et al., 2018; Kobak & Berens, 2019). 또한 선박용 디젤 엔진의 결함을 감지하는 복잡한 진동 데이터를 분류하고 시각화하기 위하여 t-SNE가 활용되기도 한다(Xi et al., 2018). Lodha & Wagh(2019)은 법률과 판례의 인용 관계를 원-핫 인코딩 방법으로 구현하고 t-SNE를 통해 2차원으로 시각화 하는 방법을 사용하는 판례의 인용 관계에 대한 분석 방법을 제안한다.

# 3. 데이터와 방법론

## 3.1. 데이터

우수 의약품 제조 기준(GMP) 규정 위반의 패턴 분석을 위하여 FDA에서 제공하는 2008년 10월부터 2022년 02월까지의 실태조사로 관찰된 기준 위반 데이터를 사용하였다. 자료는 Table 2에와 같이 3,806개의 기업을 대상으로 실시한 실태조사 과정에서 관찰된 31,218건의 규정위반 사례별로 실태조사 수행에 대한 정보(포함정보 '1', '4'), 제조소의 정보(포함정보 '2', '3', '5'), 규정 위반 내용에 대한 설명(포함정보 '6', '7', '8')으로 구성 되어있다.

**Table 2.** Type of data used for on-site inspection

| Category | Content |
|---|---|
| Data Form | 31,218rows * 8columns<br>(Regulatory violations observed in manufacturing (row)<br>* Inclusion Information (column)) |
| Inclusion Information | 1. Inspector's ID<br>2. Manufacturing plant's ID<br>3. Corporation's Name<br>4. Inspection End Date<br>5. Product Line(Medicine / Food / Cosmetic)<br>6. Violated Federal Regulations No. (CFR Number)<br>7. Brief description of the violation<br>8. Detailed description of the violation |

FDA Medical Device Databases에서 제공하는 미국연방규정(CFR) 데이터를 활용하여 211파트에 속한 293개의 완제의약품 제조소 규정 내용을 확인한다. Figure 1의 파레토 분석을 통해 총 293개의 규정 조항 중 56개의 조항에서의 위반 건수가 80.3%을 차지하고 있음을 확인하였다. Table 3는 위반 건수 상위 5개의 규정을 나타낸다.

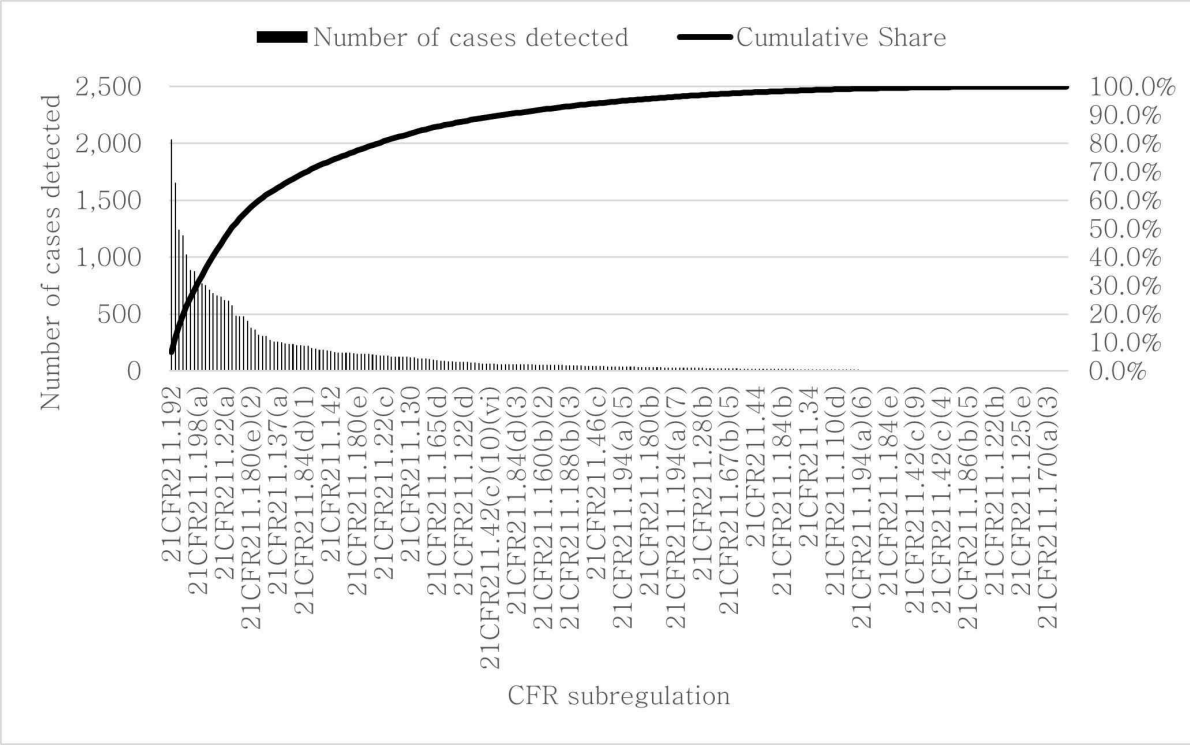


**Figure 1.** Pareto Chart of GMP Regulatory Violations

**Table 3.** Top 5 violations observed by individual regulatory units and their percentages

| No. | CFR No. | Subpart | Content | Number of violations | Percentage (%) |
|---|---|---|---|---|---|
| 1 | 21 CFR 211.192 | Subpart J. Records and Reports | Production Record Review | 2,035 | 6.5% |
| 2 | 21 CFR 211.22(d) | Subpart B. Organization and Personnel | Responsibilities of the quality unit | 1,655 | 5.3% |
| 3 | 21 CFR 211.100(a) | Subpart E. Control of Components and Drug Product Containers and Closures | Documented procedures | 1,243 | 4.0% |
| 4 | 21 CFR 211.25(a) | Subpart B. Organization and Personnel | Operator Qualification | 1,191 | 3.8% |
| 5 | 21 CFR 211.160(b) | Subpart I. Laboratory Controls | Proper sampling and testing procedures | 1,025 | 3.3% |

서브 파트 단위에서 분석한 결과 Table 4에 나타난 바와 같이 총 11개 중 10개 서브 파트에서 규정 위반이 관찰되었고 그 중 5개가 79.7%의 점유율을 보이고 있음을 확인할 수 있다.

**Table 4.** Frequency and share of violations observed on a subpart basis

| Subpart | Number of Cases detected | Percentage (%) | Cumulative Share |
|---|---|---|---|
| Subpart J. Records and Reports | 7,033 | 22.5% | 22.5% |
| Subpart I. Laboratory Controls | 5,751 | 18.4% | 40.9% |
| Subpart F Production and Process Controls | 4,525 | 14.5% | 55.4% |
| Subpart B Organization and Personnel | 4,109 | 13.2% | 68.6% |
| Subpart D Equipment | 3,452 | 11.1% | 79.7% |
| Subpart E Control of Components and Drug Product Containers and Closures | 2,109 | 6.8% | 86.5% |
| Subpart C Buildings and Facilities | 2,080 | 6.7% | 93.2% |
| Subpart G Packaging and Labeling Control | 1,345 | 4.3% | 97.5% |
| Subpart H Holding and Distribution | 699 | 2.2% | 99.7% |
| Subpart K Returned and Salvaged Drug Products | 115 | 0.4% | 100.0% |

## 3.2. 방법론

규정은 특정한 분야를 규제함에 있어서 규율 범위가 서로 필요, 충분조건의 관계를 갖는 경우가 있다. A, B, C라는 규정을 가정했을 때, 규정 준수 측면에서 C가 제시하고 있는 규정 준수의 범위가 A와 B의 내용을 내포하고 있을 경우 C는 A와 B 규정의 필요조건이 된다. [A 위반 → C 위반], [B 위반 → C 위반]으로 관찰될 가능성이 크다. 예를 들면 미국 CFR 규정 '21 CFR211.192 General requirements'의 내용은 제품이 출하되기 전에 '① 포장과 라벨링에 대한 기록을 포함하여 생산과 관련된 기록에 대해 점검을 수행'하고 '② 기록에서 모순되는 내용이 있을 경우 다른 로트(Lot)까지 확대하여 철저하게 조사'되어야 한다는 내용으로 크게 두 부분에 걸쳐 규율하고 있다. 규정 '21 CFR211.188 Batch production and control records.(b)(8)'은 제조 과정에서 기록될 내용에는 라벨링에 대한 관리 기록이 포함되어 있어야 한다는 규정으로 이를 위반할 경우 '211.192 General requirements'의 '①'의 내용을 동시에 위반하게 된다.'21 CFR211.188 Batch production and control records.(b)(12)'의 내용은 CFR211.192 General requirements에 따라 수행된 조사 내용은 기록, 관리되어야 한다는 내용의 규정으로 이를 위반할 경우 '211.192 General requirements'의 '②'의 내용을 동시에 위반하게 된다. 이는 규정 '211.192 General require-

ments'에서 제시되고 있는 준수 범위가 넓기 때문에 나타난 규율 내용의 유사성에서 비롯된 패턴으로 연관규칙 마이닝 규칙으로 도출될 것이다.

본 연구에서는 연관규칙 마이닝으로 도출된 규칙 중 규정의 내용의 상호 연관성에 기인한 패턴을 선별하기 위해 다음과 같이 네 단계로 이루어진 방법을 제안한다.

(1) 'Inspection data'에서 Apriori-Algorithm을 활용하여 빈발하는 규정 위반 집합을 형성하고 연관 분석을 통해 동시 위반 패턴을 도출한다. 내포하고 있는 의미의 복잡성을 고려하여 연관 규칙의 길이는 2로 제한한다.

(2) TF-IDF를 통하여 규정을 단어 기준으로 표현한다.

(3) TF-IDF를 통하여 고차원으로 표현된 규정을 t-SNE를 통해 2차원 공간으로 차원을 축소시킨다.

(4) 네트워크로 규정 간의 관계를 표현한다. 노드(Node)는 규정을 나타내며 위치가 가까운 노드(Node)는 규율 내용이 유사한 규정으로 해석할 수 있다. 노드(Node)의 크기는 개별 규정의 지지도(Support)와 비례하여 변화하도록 나타낸다. 동시 위반 패턴이 관찰되는 규정 간 엣지(edge)를 연결한다. 여기서 엣지(edge)의 색상은 신뢰도(Confidence)를 나타내고, 굵기는 패턴의 지지도(Support)의 값과 비례하여 나타나도록 설정한다.

# 4. 실증 분석

## 4.1. 계층적 연관 분석을 통한 규정 위반 패턴 도출

최종 단위의 항목 간의 연관규칙 분석을 진행할 때보다 상위 계층에서 연관규칙 분석을 선행하는 것은 연관규칙을 도출하는 데 계산의 효율성에 도움이 되기 때문에 연관 규칙 분석을 하향식 점진적 분석 방법으로 2차에 걸쳐 진행한다(Agrawal et al. 1993; Han and Fu, 1995). 규정의 서브 파트 수준에서 1차 연관규칙 분석을 진행한 다음 패턴이 관찰되는 서브 파트에 속하는 하위 규정에서 2차 연관규칙 분석을 진행한다.

조건부와 결과부가 동시에 나타나는 빈도 수의 비율을 나타내는 지지도(Support)의 최소값을 설정할 때 본 연구의 목적은 빈발하는 위반 규정을 찾는 것이 아닌 연관되어 나타나는 패턴을 발굴하는 것에 중점이 있기 때문에 1% 값의 낮은 기준치를 설정하였다. 조건부의 규정을 위반 했을 경우 결과부의 규정을 위반할 조건부 확률로 정의되는 신뢰도(Confidence)의 최소값은 50%로 설정하였다. 상위단계 서브파트에서 연관규칙 생성 결과 총 46개의 패턴이 도출되었다. Table 5은 서브 파트 수준에서 진행한 연관규칙 분석 결과 도출된 패턴을 신뢰도(Confidence) 순으로 정렬했을 때 상위 10개의 연관규칙을 보여준다.

**Table 5.** Regulatory violations rules derived from Subpart regulations

| No. | Antecedents | Consequents | Anteced ents support | Consequ ents support | Pattern support | con- fidence |
|---|---|---|---|---|---|---|
| 1 | Subpart G<br>Packaging and Labeling Control | Subpart J<br>Records and Reports | 0.225 | 0.682 | 0.185 | 0.823 |
| 2 | Subpart K<br>Returned and Salvaged Drug Products | Subpart J<br>Records and Reports | 0.027 | 0.682 | 0.022 | 0.814 |

| No. | Antecedents | Consequents | Anteceded ents support | Consequ ents support | Pattern support | con-fidence |
|---|---|---|---|---|---|---|
| 3 | Subpart E Control of Components and Drug Product Containers and Closures | Subpart J Records and Reports | 0.328 | 0.682 | 0.265 | 0.809 |
| 4 | Subpart E Control of Components and Drug Product Containers and Closures | Subpart I Laboratory Controls | 0.328 | 0.578 | 0.254 | 0.775 |
| 5 | Subpart C Buildings and Facilities | Subpart J Records and Reports | 0.274 | 0.682 | 0.212 | 0.774 |
| 6 | Subpart D Equipment | Subpart J Records and Reports | 0.472 | 0.682 | 0.365 | 0.774 |
| 7 | Subpart I Laboratory Controls | Subpart J Records and Reports | 0.578 | 0.682 | 0.444 | 0.768 |
| 8 | Subpart K Returned and Salvaged Drug Products | Subpart B Organization and Personnel | 0.027 | 0.597 | 0.020 | 0.765 |
| 9 | Subpart G Packaging and Labeling Control | Subpart B Organization and Personnel | 0.225 | 0.597 | 0.171 | 0.761 |
| 10 | Subpart F Production and Process Controls | Subpart J Records and Reports | 0.547 | 0.682 | 0.415 | 0.760 |

서브 파트에서의 패턴보다 하위 규정 간의 패턴이 보다 구체적인 사례를 제시할 수 있다. 이에 서브 파트의 연관규칙 분석 결과를 신뢰도(Confidence) 순으로 정렬했을 때 상위 10개의 패턴에 속하는 하위 규정 간의 연관규칙 분석을 진행하였다. Table 6은 서브 파트 하위 규정 간 연관 규칙 분석에서 관찰된 패턴의 수를 나타낸다.

**Table 6.** Association rules between subpart, sub-regulations

| No. | Antecedents | Consequents | Number of patterns observed in sub regulation |
|---|---|---|---|
| 1 | Subpart G Packaging and Labeling Control | Subpart J Records and Reports | 11 |
| 2 | Subpart K Returned and Salvaged Drug Products | Subpart J Records and Reports | 11 |
| 3 | Subpart E Control of Components and Drug Product Containers and Closures | Subpart J Records and Reports | 14 |
| 4 | Subpart E Control of Components and Drug Product Containers and Closures | Subpart I Laboratory Controls | 2 |

| No. | Antecedents | Consequents | Number of patterns observed in sub regulation |
|---|---|---|---|
| 5 | Subpart C<br>Buildings and Facilities | Subpart J<br>Records and Reports | 25 |
| 6 | Subpart D<br>Equipment | Subpart J<br>Records and Reports | 14 |
| 7 | Subpart I<br>Laboratory Controls | Subpart J<br>Records and Reports | 24 |
| 8 | Subpart K<br>Returned and Salvaged Drug Products | Subpart B<br>Organization and Personnel | 1 |
| 9 | Subpart G<br>Packaging and Labeling Control | Subpart B<br>Organization and Personnel | 3 |
| 10 | Subpart F<br>Production and Process Controls | Subpart J<br>Records and Reports | 22 |

## 4.2. TF-IDF를 통한 규정 텍스트 정보의 벡터 변환

본 연구에서는 Table 3의 293개 하위 규정 내용의 유사성을 확인하기 위해 TF-IDF를 활용하여 각 규정 텍스트를 단어 기준으로 표현한다. 규정 간 준용을 위해 사용된 숫자와 특별한 의미를 갖지 않는 전치사, 부사 형태는 불용어 처리하였다. 또한 출현하는 문서 수가 1개로 산출되는 단어는 제외하였다. 명사의 단수, 복수 형태를 나타내거나 주어의 인칭 차이에 의해 '-s'가 붙는 단어들은 동의어 처리를 진행하였고 어간 추출을 위해 Porter 2 알고리즘을 사용하였다. 분석 결과 293개의 규정은 624개의 단어로 표현되었다. Table 7은 단어의 빈도 분석(Term Frequency) 결과 상위 10개의 단어와 출현 수를 나타낸다.

**Table 7.** Frequency Value of Top 10 words

| No. | Word | Frequency |
|---|---|---|
| 1 | product | 557 |
| 2 | drug | 442 |
| 3 | control | 201 |
| 4 | record | 198 |
| 5 | test | 153 |
| 6 | batch | 152 |
| 7 | sampl | 134 |
| 8 | procedur | 127 |
| 9 | use | 124 |
| 10 | written | 123 |

## 4.3. t-SNE를 이용한 차원 축소

t-SNE는 차원 축소 시 TF-IDF로 고차원의 벡터로 표현된 규정 위치 거리 간의 확률적 유사성을 Gaussian Kernel을 통해 나타내고 규정 간의 관계를 시각화하기 위한 2차원의 공간에서의 데이터 거리 간의 확률적 유사성은 자유도 1의 Student-t 분포로 나타내어 각 차원의 데이터 위치 유사성을 Kullback-Leibler divergences 측면에서 가능한 동일한 값을 갖도록 데이터를 위치시킨다. t-SNE의 파라미터 perplexity는 사용자가 입력하는 값으로 TF-IDF로 고차원의 벡터로 표현된 규정에서 내용의 유사성에 의해 군집을 이루는 규정의 개수로 해석할 수 있다 (Van & Hinton, 2008). Wattenberg et al.(2016)는 연구에서 perplexity을 5~50의 값으로 설정할 때 t-SNE는 강건한 성능을 나타냄을 보였다. 본 연구에서 사용하는 데이터는 '규정'으로 내용적으로 연관된 규정이 많지 않은 특성을 갖고 있음으로 파라미터 perplexity가 강건한 성능을 나타내는 범위에서 가장 낮은 값인 5로 설정하였다.

축소된 차원에서 규정을 나타내는 노드(Node) 위치가 규율하고 있는 내용이 유사할 경우 거리가 가깝게 위치하도록 분포하는데 t-SNE을 사용한 차원 축소가 적절 하였는지 확인하기 위해 TF-IDF를 통해 벡터로 표현된 293개의 규정을 t-SNE, PCA(Principal component analysis), LLE(Locally Linear Embedding), MDS(Multi Dimensional Scailing)를 사용하여 차원 축소하고 결과를 비교하였다. 규정의 내용 확인을 통해 규율하고 있는 내용이 유사한 28개 규정을 Table 8에 나타냈다. 제시된 28개 규정은 가깝게 위치하고 나머지 규정은 멀게 위치해야 차원 축소가 적절하게 이루어진 것으로 판단할 수 있다. Table 9는 각 모형에서 설정된 파라미터 값을 나타낸다. Figure 2는 각 방법을 통해 차원 축소된 결과를 나타낸다. PCA를 사용한 방법(b)은 유사한 규정 간의 거리를 가깝게 표현하지 못한다. LLE와 MDS를 활용한 차원 축소 방법(c, d)에서는 내용적으로 유사한 규정과 유사하지 않은 규정 간 거리의 분명한 차이가 관찰되지 않는다. 반면에 t-SNE를 통한 차원축소 방법(a)에서는 규제 내용이 비슷한 규정이 밀집하여 군집을 이루고 있고, 다른 규정은 상대적으로 먼 위치에 있음을 확인할 수 있다. t-SNE를 적용한 차원 축소 방법이 축소된 차원에서의 규정 간의 필요, 충분조건 관계를 가장 우수하게 나타낸다.

**Table 8.** 28 regulations with similar scope of discipline

| Regulation | Short Description | Characteristics of node representation |
|---|---|---|
| 21CFR211.192 | Regulate the obligation to review production and control records before distribution and to conduct an investigation on deviations, etc. | Color : red |
| 11 regulations including 21CFR211.186(a) ~ 21CFR211.186(b)(9) | Regulate the need for master production and control records and what contents to be included. | Color : green |
| 16 regulations including 21CFR211.188 ~ 21CFR211.188(b)(13) | Regulate the obligation to prepare all batch's production and control records and what contents to be included. | Color : blue |

**Table 9.** Parameter setting values for the experiments

| Technique | Parameters | Setting value |
|---|---|---|
| t-SNE | perplexity | 5 |
| PCA | none | - |
| LLE | neighbors | 3 |
| MDS | none | - |

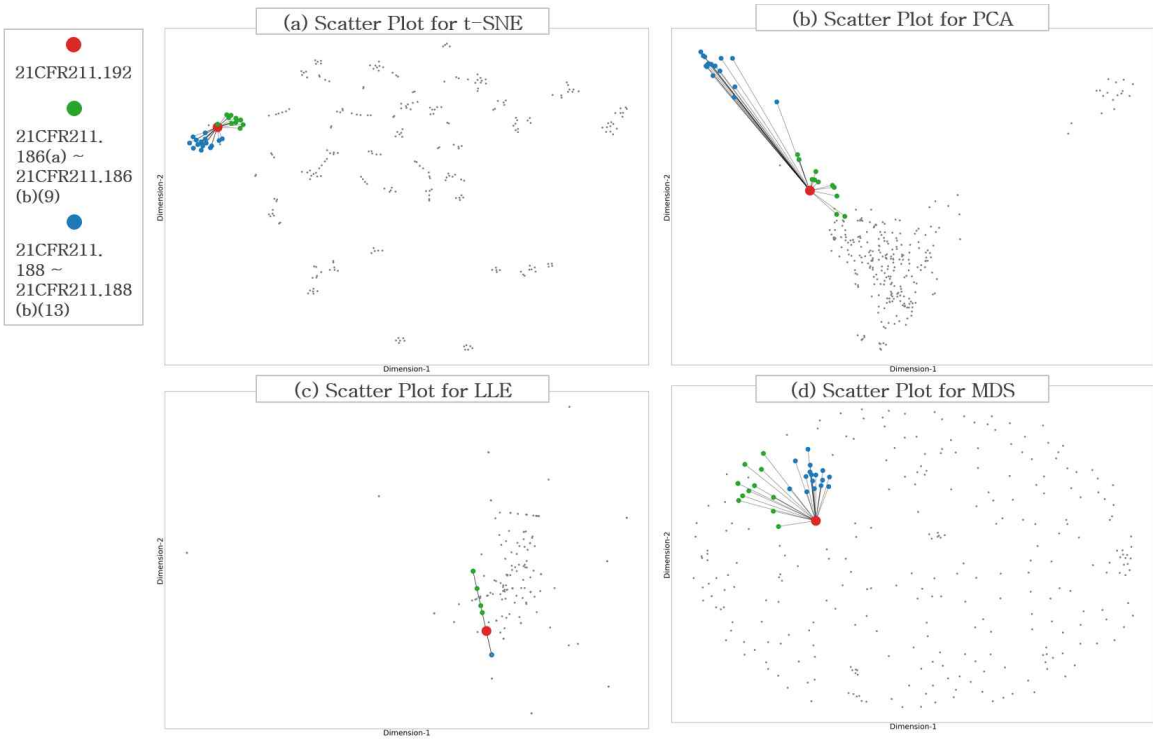


**Figure 2.** Dimension reduction visualization of vectorized regulation using t-SNE, PCA, LLE, MDS

## 4.4. 네트워크 표현

t-SNE를 통해 2차원으로 축소된 데이터를 네트워크 형태로 나타낸다. Figure 3는 Table 5에서 가장 큰 신뢰도(Confidence) 값이 관찰된 '서브파트 G 포장 및 라벨링 관리'와 '서브파트 J 기록 및 보고서'의 하위 규정 간의 패턴을 네트워크 형태로 나타낸 것이다. 하위 규정을 나타내는 노드(Node)의 크기는 개별 하위 규정의 지지도(Support)와 비례하여 변화하도록 나타냈다. 연관규칙 패턴이 관찰되는 규정 간 엣지(edge)를 연결한다. 여기서 엣지(edge)의 색상 진한 정도를 신뢰도(Confidence)의 크기로 나타내고, 굵기는 패턴의 지지도(Support)의 값을 나타내도록 설정하였다. 패턴이 관찰된 규정 노드(Node) 사이의 유클리드 거리(Euclidean Distance)의 값과 규정 간 나타나는 모든 유클리드 거리의 평균 값의 차이를 표준편차로 나눈 표준화 값을 사용하였다. 규제하고 있는 내용이 상호 연관된 규정 사이에서 관찰되는 노드(Node)의 유클리드 거리 값을 분석하여 노드 간 거리가 -1.0 이하의 표준화 값으로 관측

되는 패턴은 규율 내용의 유사성에 기인하여 나타나는 규칙으로 판단하였다.

Figure 4는 규정 21 CFR211.192를 기준으로 다른 규정과의 거리 표준화 값을 등고선으로 나타낸 것이다. 21 CFR211.188(b)(8)은 노드 간 거리의 표준화 값 -1.0 이하에서 관찰된다. Table 10에 나타난 바와 같이 21 CFR211.188(b)(8)은 라벨링 관리와 관련하여 제조기록서에 담겨야 하는 내용을 규제하고 있다. 21 CFR211.192은 출하 전 모든 제조 관련 기록서는 검토되야 함을 규율하고 있다. 두 개의 규정은 규제하고 있는 내용이 높은 상호 연관성이 있어 규율 내용의 유사성에 의해 나타난 패턴으로 해석할 수 있다. 반면 표준화 값 -1.0의 외부에 위치한 CFR211.122(d)은 Table 10에 나타난 바와 같이 다른 제품에 사용되는 라벨 등의 자재는 식별이 용이한 상태로 보관해야 한다는 것을 규제하고 있어 규정 21 CFR211.192와 유사성이 없는 규정에서 발견되는 동시 위반 패턴으로 판단할 수 있다.

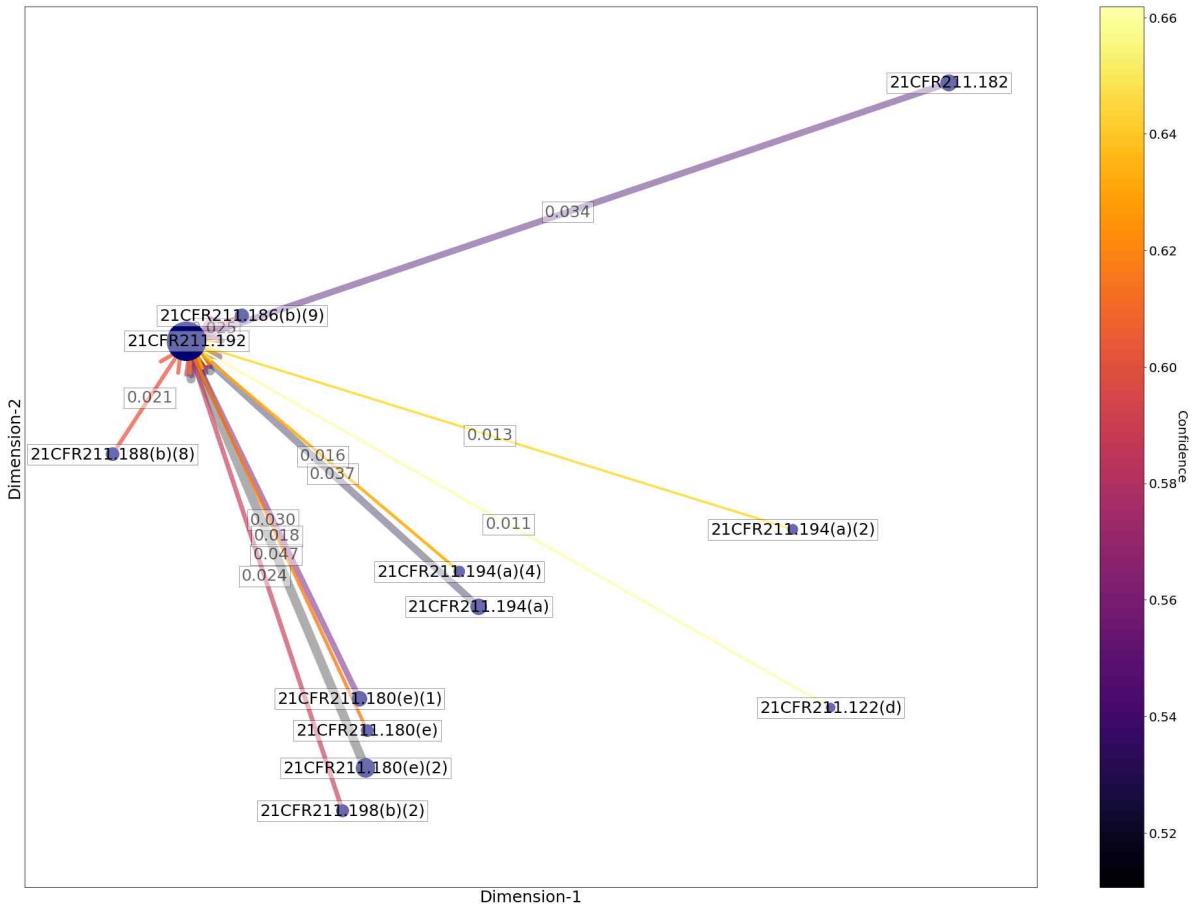


**Figure 3.** Network graph for analysis of violation patterns in sub-regulation between Subpart G 'Packaging and Labeling Control' and Subpart J 'Records and Reports'

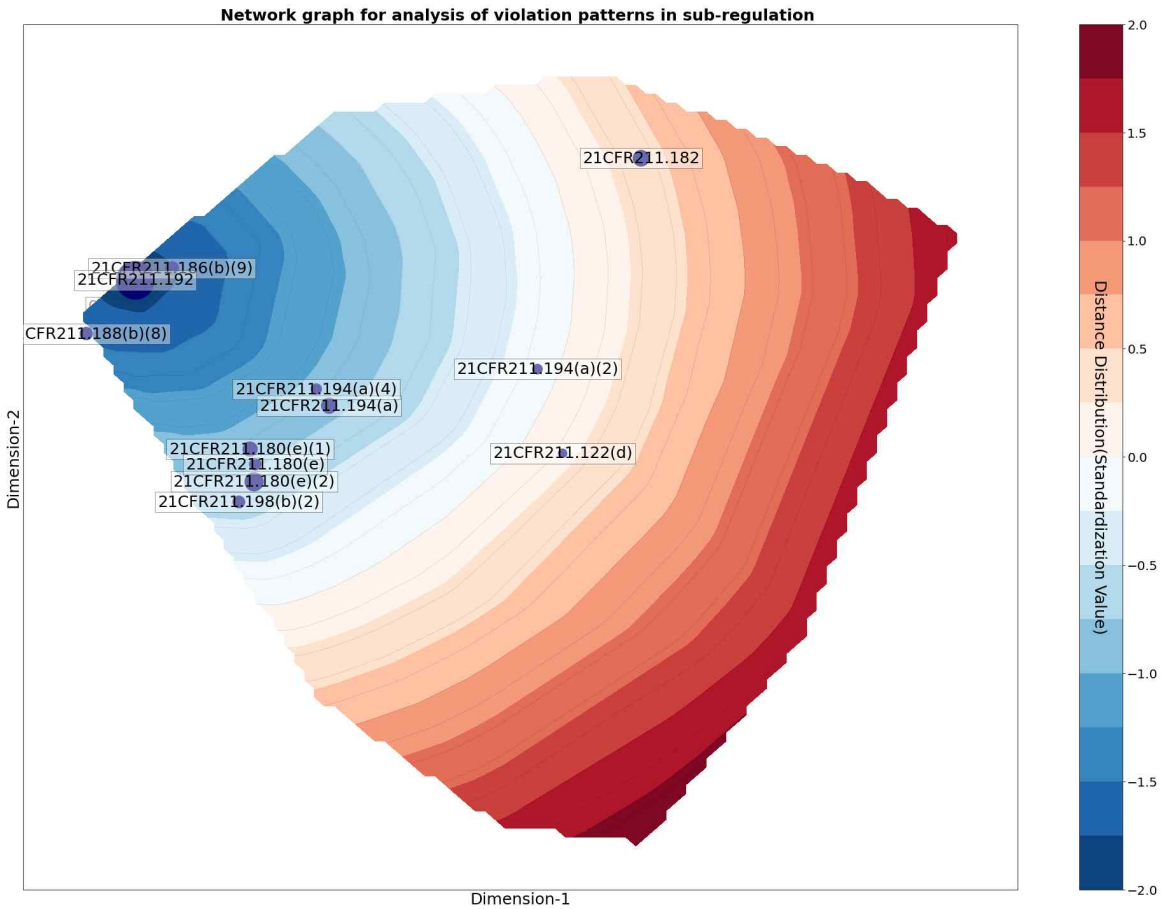


**Figure 4.** Contour plot showing the standardized values of distance between nodes in a network graph for analysis of sub-regulation violation patterns between Subpart G 'Packaging and Labeling Control' and Subpart J 'Recordings and Reports'

**Table 10.** Content of regulations [21 CFR 211.122(d), 21 CFR 211.188(b)(8), 21 CFR211.192] and location characteristics between two nodes shown in Figure 3, 4

| Antecedents | Consequents | Confidence | Antecedents sub-regulation content | Consequents Sub-regulation content | Standardization Value (Close : Standardization Value ≤ -1.0) |
|---|---|---|---|---|---|
| 21 CFR 211.188 (b)(8) | 21 CFR 211.192 | 0.519 | Batch production and control records shall be prepared for each batch of drug product produced and shall include complete information relating to the production and control of each batch. These records shall | All drug product production and control records, including those for packaging and labeling, shall be reviewed and approved by the quality control | Standardization Value = -1.521 (Close) |

| Antecedents | Consequents | Confidence | Antecedents sub-regulation content | Consequents Sub-regulation content | Standardization Value (Close : Standardization Value ≤ -1.0) |
|---|---|---|---|---|---|
| | | | include: Documentation that each significant step in the manufacture, processing, packing, or holding of the batch was accomplished, including: Complete labeling control records, including specimens or copies of all labeling used;. | unit to determine compliance with all established, approved written procedures before a batch is released or distributed. (…) | |
| 21 CFR 211.122 (d) | 21 CFR 211.192 | 0.540 | Labels and other labeling materials for each different drug product, strength, dosage form, or quantity of contents shall be stored separately with suitable identification. Access to the storage area shall be limited to authorized personnel. | ↑ | Standardization Value = 0.363 (Not Close) |

# 5. 연구의 결론과 시사점

본 연구에서는 실태조사가 갖는 감사 자원의 투입 한계와 제조소의 은닉 규정 위반 실태 적발의 어려움을 극복하고자 동시에 나타나는 규정 위반 패턴을 분석하기 위한 계층적 연관규칙과 텍스트 마이닝 기반 t-SNE 분석 방법을 제시하였다. 기존의 실태조사 효과성을 재고하기 위한 연구는 주로 개별 규정의 위반 경향과 최종 GMP 적합 판정율에 영향을 미치는 요인을 탐색하는 것과 집중해서 조사해야 할 감사 분야 선정 방법 중심으로 이루어져 왔다. 이러한 연구는 특정 분야의 실정을 면밀하게 관찰하는데 도움이 될 수는 있지만, 선정되지 않은 분야에서의 실태 조사는 소홀하게 이루어질 수 있는 위험이 있다. 또한 실태조사 중에 적발된 위반 사례를 토대로 감추어진 위반 실태를 적발해 내거나, 적발하는데 소요되는 시간을 단축하는 것에도 기여하지 못한다.

동시에 나타나는 규정 위반 패턴 인식 방법에 대한 연구는 어떠한 위반 사례가 적발되었을 경우 동시적으로 위반할 가능성이 높은 규정을 감사 분야에 제약하지 않고 제시할 수 있도록 한다. 또한 의도적으로 은폐된 위반을 발견할 가능성을 높이고 적발 소요 시간 단축에도 기여할 것이다. 의약품 제조소 실태조사 과정에 본 연구에서 제시하는 방법을 활용한다면 제조소의 우수 의약품 제조 기준(GMP) 준수 실태를 보다 더 잘 살펴볼 수 있을 것이며 나아가 의약품 제조사가 적법한 절차를 준수하고 사용 목적에 부합하는 양질의 의약품을 지속적으로 제조, 공급하도록 함으로써

공중 보건의 질 향상에 기여할 수 있을 것이다.

미국과 유럽에서는 의약품 제조소 실태조사 결과에 대한 정보가 축적되고 공개되고 있으며, 이를 활용한 연구가 진행되고 있어 제조공장의 품질향상과 효과적인 실태조사가 이루어지는데 기여하고 있다. 최근 우리나라의 식품의약품안전처는 국내 의약품 제조소의 신뢰도 제고를 위하여 GMP 실태조사 결과를 대중에게 공개하는 것을 추진하고 있다. 본 연구에서 제안된 방법은 앞으로 공개될 국내 제약사들의 실태조사 자료에 충분히 활용될 수 있을 것으로 기대된다.

아울러 규정 위반 데이터를 대상으로 한 연관규칙 분석 결과로써 나타나는 수많은 패턴들 중에서 서로 다른 규정의 내용이 유사하거나 상호 연관성이 높아 나타나는 패턴을 빠르게 식별할 수 있도록 하는 방법을 제시했다는 점도 의미가 있다. 이 방법은 본 연구에서 수행된 의약품 제조소 감사 결과 자료에 한정되지 않고 다양한 분야의 감사 자료 분석에 응용될 수 있을 것이다.

본 연구에서 활용된 t-SNE를 통한 차원 축소 방법은 두 가지의 한계가 있다. 첫 번째 한계는 축소된 차원의 특정 위치에 표현되고 있는 규정은 어떤 요인을 원인으로 유사한 규정으로 판단되어 가깝게 위치하는지에 대해서 명확히 설명되지 않는 것이다. 두 번째는 네트워크에서 표현된 노드(Node)간 거리가 얼마나 멀어야 규정의 규율 범위의 상호 유사성이 없는지 판단이 어렵다는 것이다. 이러한 점을 개선하기 위해 규정 간의 규율 유사성 여부를 판단할 수 있는 노드(Node) 간의 거리 차이를 명확히 제시하면서 축소된 차원에서 노드(Node)의 위치를 결정하는 요인을 설명할 수 있는 방법에 대한 분석을 향후 과제로 남겨놓고 있다.

# REFERENCES

Abdellah, A., Noordin M., Zaki R., Abedallah, A. 2016. Pharmaceutical good manufacturing practice regulatory affairs in Sudan : continuous debate between regulatory authority and manufacturers. Pharmaceutical Regulatory Affairs 5(1):166.

Agrawal, R., Imieliński, T., Swami, A. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of data, pp. 207-216.

Ban, E., Jang, D., Kwon, T., Kim, A. 2019. Pharmaceutical GMP inspection: case study of USA and EU. Yakhak Hoeji 63(1):15-23.

Choi, G., & Park, J. 2018. The Effects of Audit Quality on Audit Performance. Journal of Korean Society for Quality Management 46(1):113-134.

Han, J., Fu, Y. 1995. Discovery of multiple-level association rules from large databases. In Proceedings of 1995 International Conference on Very Large Data Bases 95: 420-431.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G., Chen, M., Guo, G. 2018. Mapping the mouse cell atlas by microwell-seq. Cell 172(5):1091-1107.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). 2008. PHARMACEUTICAL QUALITY SYSTEM, (ICH Publication Code. Q10). Geneva : ICH

Kim, S., Park, E., Cho, H., Hong, S., Sohn, B., & Hong, J. 2021. Pattern Analysis of Nonconforming Farmers in Residual Pesticides using Exploratory Data Analysis and Association Rule Analysis. Journal of Korean Society for Quality Management 49(1):81-95.

Kobak, D., Berens, P. 2019. The art of using t-SNE for single-cell transcriptomics. Nature Communications 10(1): 1-14.

Kumar, S., Tanwar, D., Arora, N. 2013. The role of regulatory GMP audit in pharmaceutical companies. International Journal of Research and Development in Pharmacy and Life Sciences 2(4):493-498.

Lee, S. B., Kang, G., Jang, W., & Kwon, B. J. 2022. A Study on the Development and Effectiveness of ISO13485: 2016 Education Programs for Domestic Medical Device Manufacturers. Journal of Korean Society for Quality Management 50(3):503-515.

Lim, J., Cho, H. 2021. An Analysis of the Factors Related to EU GMP Non-compliance-Focusing on Active Substance, Human Medicinal Product. Yakhak Hoeji 65(1):46-55.

Lin, W., Alvarez, S. A., Ruiz, C. 2002. Efficient adaptive-support association rule mining for recommender systems. Data Mining and Knowledge Discovery 6(1):83-105.

Linna A, Korhonen M, Mannermaa JP, Airaksinen M, Juppo AM. Developing a tool for the preparation of GMP audit of pharmaceutical contract manufacturer. European Journal of Pharmaceutics and Biopharmaceutics. 2008 Jun;69(2):786-92.

Lodha, S., Wagh, R. 2019. Exploratory analysis of legal case citation data using node embedding. ICIC Express Letters 13(10):883-889.

Onan, A., Bulut, H., Korukoglu, S. 2017. An improved ant algorithm with LDA-based representation for text document clustering. Journal of Information Science 43(2):275-292

P. Bafna., D. Pramod., A. Vaidya. 2016. Document clustering: TF-IDF approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 61-66.

Pharmaceutical Inspection Convention Pharmaceutical Inspection Co-operation Scheme (PIC/S). 2009. Quality System Requirements for Pharmaceutical Inspectorates, (PIC/S Publication No. PI002-3). Geneva : PIC/S

Pharmaceutical Inspection Convention Pharmaceutical Inspection Co-operation Scheme (PIC/S). 2009. Standard operating procedure Team inspections, (PIC/S Publication No. PI031-1). Geneva : PIC/S.

Pharmaceutical Inspection Convention Pharmaceutical Inspection Co-operation Scheme (PIC/S). 2012. A recommended model for risk-based inspection planning in the GMP environment, (PIC/S Publication No. PI0371). Geneva : PIC/S.

Seong, Y., Lee, D., Jung, U. 2022. Group-wise Keyword Extraction of the External Audit using Text Mining and Association Rules. Journal of Korean Society for Quality Management 50(1):77-89.

Tao, F., Murtagh, F., Farid, M. 2003. Weighted association rule mining using weighted support and significance framework. In Proceedings of the ninth ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 661-666.

Van der Maaten, L., Hinton, G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11).

Waldron, K. 2017. Managing risk to the patient: recoding quality risk management for the pharmaceutical and biopharmaceutical industries. Doctoral dissertation, Dublin Institute of Technology, Ireland.

Wang, W., Yang, J., Yu, P. 2004. WAR: weighted association rules for Item intensities. Knowledge and Information Systems 6(2):203-229.

Wattenberg, M., Viégas, F., Johnson, I. 2016. How to use t-SNE effectively. Distill 1(10):e2.

Xi, W., Li, Z., Tian, Z., Duan, Z. 2018. A feature extraction and visualization method for fault detection of marine diesel engines. Measurement 116:429-437.

## 저자소개

**이준오** 현재 연세대학교 산업공학과 석사과정 재학중이다. 주요 관심분야는 품질공학, 데이터 마이닝 연구이다.

**손소영** 연세대학교 수학과를 졸업한 후 과학원 산업공학과와 임페리얼대학교 경영과학과에서 석사 과정을 마치고, 피츠버그대학교에서 응용통계학석사, 산업공학박사 학위를 받았다. 현재 연세대학교 산업공학과 교수이다. 주요 관심분야는 품질공학, 특허경영전략 과학화, 공간 빅데이터 마이닝 연구이다.