Korean J Otorhinolaryngol-Head Neck Surg / eISSN 2092-6529
https://doi.org/10.3342/kjorl-hns.2025.00122

Design and Development of Medical Databases

Tae Hoon Kong[1,2,3,4] and **Young Joon Seo**[1,2,3,4]

Departments of [1]Otorhinolaryngology-Head and Neck Surgery, [2]Medical Informatics and Biostatistics, and [3]Bio-AI Convergence, Yonsei University Wonju College of Medicine, Wonju; and
[4]Research Institute of Hearing Enhancement, Yonsei University Wonju College of Medicine, Wonju, Korea

의학 데이터베이스의 설계와 구축

공태훈[1,2,3,4] · **서영준**[1,2,3,4]

연세대학교 원주의과대학 [1]이비인후과학교실, [2]의료정보통계학과, [3]바이오-AI 융합학과, [4]청각재활연구소

Since the Fourth Industrial Revolution was declared at the World Economic Forum in 2016, we have witnessed significant technological innovations across various fields. Artificial intelligence (AI) technology has developed alongside computing technology, algorithms, and platforms, but its most crucial foundation lies in data collection and management capabilities. This paper aims to share expertise gained through our experience in constructing medical databases, including the Hearing Big Data Center, Common Data Model implementation, and balance function test databases. We first explain the fundamental concepts of databases, distinguishing between structured and unstructured data, and introduce relational database models and data schemas particularly relevant to otolaryngology. We then describe practical strategies for designing and operating medical databases that can flexibly handle both structured and unstructured data, using hearing and vestibular function test data as representative examples. Additionally, we address important considerations for healthcare database construction, including privacy regulations, de-identification processes, and data combination techniques according to South Korean guidelines. Finally, we discuss how well-designed medical databases can be linked with AI and machine learning models, enabling multimodal analysis that integrates clinical data, imaging, and physiological signals. We argue that by understanding database design principles and proper data management approaches, medical researchers can more efficiently utilize vast medical data resources while maintaining patient privacy, ultimately advancing medical research and clinical practice in the era of AI.

Korean J Otorhinolaryngol-Head Neck Surg

Keywords Artificial intelligence; Databases; Data collection; De-identification; Otolaryngology.

Received February 26, 2025
Revised December 10, 2025
Accepted December 15, 2025

Address for correspondence
Young Joon Seo, MD, PhD
Department of Otorhinolaryngology-Head and Neck Surgery,
Yonsei University Wonju College of Medicine,
20 Ilsan-ro, Wonju 26426, Korea
Tel +82-33-741-0642
Fax +82-33-732-8287
E-mail okas2000@yonsei.ac.kr

서 론

2016년 세계경제포럼(World Economic Forum, WEF)에서 4차 산업혁명 시대를 선언한 이후 우리는 실제로 4차 산업혁명의 핵심인 빅데이터, 인공지능, 로봇공학, 사물인터넷, 무인항공기, 3차원 프린팅, 나노기술 등 4차 산업혁명의 6대 중점분야의 기술 혁신을 어렵지 않게 접할 수 있게 되었다.[1,2] 2016년 구글 딥마인드는 '알파고'를 통하여, 그동안 상당한 창의력이 필요해 인간의 영역이라고 보았던 바둑에서 인간계의 최고인 이세돌을 상대로 4승 1패의 성적을 거두면서 뛰어난 인공지능의 실력을 보여주었고, 테슬라는 전기차를 통하여 자동차 산업의 혁신을 이루어 냈다고 평가받지만, 그보다 크게 조명 받는 것은 전기차를 통해 보여주는 인공지능을 활

용한 자율주행 기술이다. 생성형 인공지능의 대표주자들이 대중에 알려지게 되면서 이제 ChatGPT를 포함한 거대언어 모델 인공지능을 사용하며 마치 개인 비서를 고용한듯 업무의 효율을 높이게 되었다.[3,4]

인공지능 기술은 컴퓨터 기술의 발달과 더불어 알고리즘 및 플랫폼의 발달에 배경이 있으나, 더욱 중요한 것은 방대한 데이터를 수집하고 관리하는 기술의 발달에 그 배경이 있다. 컴퓨터 기술의 발달과 클라우드 컴퓨팅은 방대한 데이터의 저장과 관리, 조회뿐 아니라 실시간 분석과 시각화와 같은 데이터 활용의 기회를 제공한다.[5] 이러한 흐름은 빅데이터의 소위 5V (Variety [다양성], Volume [규모], Velocity [속도], 가치 [Value] 혹은 정합성 [Veracity])로 일컬어지며 빅데이터는 단순한 대형 데이터뿐 아니라 텍스트, 로그(log), 이미지 등 광범위한 분야에서 발생하는 각양각색의 데이터를 모아 활용하겠다는 의지에서 출발한다.[6]

의학에서도 특히 이비인후과에서 사용하는 데이터는 청력검사, 평형기능검사, 방사선·MRI 영상, 전자의무기록(electronic medical record) 등 비정형화된 경우가 많다. 이로 인해 의학연구에서 뿐 아니라 진료 전반에서의 데이터를 수집, 활용 및 관리하는 과정에서 병원과 장비, 검사 프로토콜에 따라 구조와 포맷이 제각각이고, 장기적인 연구·AI 개발 관점에서 재사용성이 떨어지는 문제가 있다. 이러한 배경에서 저자들은 대한청각학회 청각빅데이터센터 구축 사업, 공통데이터모델(Common Data Model, CDM) 구축 사업, 어지럼증과 관련된 평형기능검사 데이터베이스 구축을 수행하며, 이비인후과 임상 현장에 적합한 데이터베이스 설계와 운용에 대한 실질적인 노하우를 축적하게 되었다. 본 논문에서는 이 경험을 바탕으로, 정형·비정형 데이터를 동시에 다루는 이비인후과 데이터베이스 설계 원칙과 구체적인 구현 사례를 소개하고자 한다.

데이터베이스의 이해: 정형·비정형 데이터

데이터베이스의 사전적 의미는 '양적, 질적인 변수 형태의 값으로 이루어진 것들'로 정의된다.[7] 물론 빅데이터 시대에는 양적, 질적이라고 정의하기 어려운 형태의 텍스트, 음성, 이미지, 영상 등 모두 데이터의 범주에 들어간다. 그러나 데이터 분석을 위해서는 사전적 의미의 데이터의 형태로 정리되어야 하며, 이를 위해서는 정형데이터와 비정형데이터에 대한 이해가 필요하다.

정형데이터는 미리 정의된 데이터의 구조 또는 모형을 따르는 데이터로 사전에 정의된 체계적인 시스템(데이터 스키마, data schema)으로 구성되며 양적 특성을 받으므로 데이터의 검색과 분석이 매우 쉽다. 정형데이터는 일반적으로 데이터웨어하우스에 저장되며 structured query language (SQL) 기반의 관계형 데이터베이스가 사용된다.[7]

비정형 데이터는 정형데이터를 제외한 나머지를 모두 일컫는 것으로 생각하면 쉽고 텍스트, 음성, 이미지, 영상 등 모든 형태로 존재할 수 있다. 이비인후과 영역뿐 아니라 의학에서 사용되고 생성되는 각종 검사 결과지, 방사선 혹은 MRI 영상, 전자의무기록의 형태로 기록되는 진료기록 등은 대부분 비정형데이터의 범주에 포함된다. 비정형데이터는 특정 데이터 스키마가 포함되지 않고 유연한 형태를 가지지만 데이터의 특성상 검색과 분석이 어렵다는 단점이 있다.[8] 따라서 이비인후과에서 발생하는 다양한 형태의 정형·비정형 데이터를 통합적으로 다루기 위해서는, 두 유형의 데이터를 각각 어떻게 구조화하고 연결할 것인지에 대한 설계가 필수적이다.[8]

관계형 데이터베이스 구조와 데이터 스키마

1970년 Codd[9]가 처음 제안한 뒤로 현재까지 널리 사용되고 있는 데이터베이스 모델은 관계형 데이터베이스 모델이다.[9] 관계형 데이터베이스 모델은 하나의 개체가 가지고 있는 무수히 많은 속성이나 특성에 대하여 각각의 개념별로 데이터를 관계 대수와 관계 해석에 의해서 데이터 간의 정보를 2차원 테이블 형태로 표현하는 것을 말한다. Fig. 1은 관계형 데이터베이스 모델의 예시로 '학생'과 '교수', 그리고 '개설과목' 개체의 관계를 데이터베이스 테이블에 표시한 예를 보여준다. 각 테이블에서 '학생' 개체와 '교수' 개체 사이에는 '지도'라는 관계가 있고, '교수번호'라는 속성이 서로를 연결해주는 고리로 이용되고 있다. 또한 '학생', '교수', '과목' 개체 사이에는 강좌의 개설과 수강의 관계가 있으며 '개설학과' 속성이 서로를 연결해주는 고리로 이용되고 있다. 이렇듯 각 개체가 가지고 있는 무수히 많은 속성이나 특성을 각각의 개념별로 테이블을 생성하여 데이터베이스를 구성하고 정리한 것을 관계형 데이터베이스라고 한다.

데이터베이스의 구조와 제약 조건에 대하여 전반적인 명세를 기술한 것을 데이터 스키마라고 한다. 데이터 스키마는 데이터베이스에서 특성을 나타내는 속성(attribute)과 속성들의 집합으로 이루어진 개체(entity), 개체 사이에 존재하는 관계(relation)에 대한 정의와 이들을 유지해야 할 제약조건을 기술한 것을 말한다. 관계형 데이터베이스에는 필수적으로 관계형 스키마를 가지게 되는데 관계형 데이터베이스에서 관계형 스키마란 위에서 설명한 Fig. 1 예시에서 제시한 것과 같이

A

Student_ID	Student_Name	Department	Grade	Professor_code
2023122445	Hong G	Statistics	2	1234
2021223224	Kim C	Computer	3	4072
2023525657	Kim Y	Statistics	3	5456
2022565789	Lee S	Informatics	2	6668
2024333578	Kim S	Computer	2	8723

B

Professor_code	Professor_Name	Department
1234	Kong	Statistics
4077	Seo	Computer
5456	Kim	Statistics
6668	Lee	Informatics
8723	Park	Computer

C

Subject_No	Subject_Name	Department
AA1234	Math1	Statistics
AA5678	Medicine	Informatics
AA2345	English	Common
AB3456	Math2	Statistics
AC9876	Data1	Computer

Fig. 1. Example of a Relational Database Model. This figure shows an example of a relational database model displaying the relationships between "Student," "Professor," and "Course Offering" entities in database tables. The relationship "Guidance" connects the "Student" and "Professor" entities through the "Professor ID" attribute. Additionally, relationships of course offerings and registrations exist among the "Student," "Professor," and "Course" entities, with the "Department" attribute serving as the connecting link. This illustrates how relational databases organize attributes of each entity into concept-based tables with defined relationships. A: Student Table. B: Professor Table. C: Subject Table.

'학생'과 '교수' 그리고 '개설과목'의 관계형 데이터베이스에서 각 개체와 관계를 정의하는 것을 말하는 것으로, 데이터베이스 설계에서 필수적인 요소라고 할 수 있다.[7]

의료 데이터 베이스 구축: 이비인후과 코호트와 하이브리드 구조

대부분의 연구자들이 본인의 연구 가설을 검증하기 위해 데이터를 구축하고 분석하는데, 특정 가설 하나만을 검증하기 위해 데이터를 구축하는 것은 매우 비효율적인 일이다. 따라서 많은 연구자들이 관심 있는 질환에 대한 전향적 혹은 후향적 코호트를 구축하고 이를 활용하는 방법을 많이 사용하며, 이때 구축된 코호트 또한 하나의 데이터베이스로 간주될 수 있다. 코호트 데이터는 연구계획당시 설계된 증례기록서(case report form)대로 엑셀 파일 등의 형태로 저장되는 경우가 많으나 데이터베이스를 이렇게 생성하고 보관하는 것은 확장성을 고려한 데이터베이스 관점에서는 비효율적이라는 지적이 있어 왔다.[10]

저자들이 2020년 대한청각학회 주관 청각빅데이터 구축 사업 당시 설계한 관계형 데이터베이스와 데이터 스키마는 간단한 구조로 되어 있지만 확장성 측면에서 큰 의미가 있다(Fig. 2). 청각빅데이터센터 구축 당시 약 12만 건의 순음청력검사(pure-tone audiometry, PTA), 8천여 건의 어음청력검사, 수만 건의 진단명, 메타데이터가 결합되는 복합 구조의 데이터베이스를 설계하였다. 특히 기존 청력검사지는 대부분 병원별로 형태가 상이하고 구조화되어 있지 않아, 데이터를 그대로 저장하면 SQL 기반 데이터 분석이 어렵다는 문제점이 있었다. 저자들은 이를 수기로 수치화함과 동시에 환자 테이블과 검사 테이블을 1:N 구조로 연결하고, 각 주파수별 역치는 long-form 구조로 저장하여 데이터 확장성과 SQL 기반 분석의 효율성을 확보하였다. 이 방식은 향후 새로운 주파수 추가(예: extended high frequency), bone/air conduction 추가, masking 여부 추가 등 확장성 있는 구조를 보장할 뿐 아니라 SQL 조건문을 통한 통합 분석 효율성을 극대화한다.

또한 최근 저자들이 자체적으로 구축한 전정기능검사 데이터의 경우 비정형 데이터를 포함하기 때문에 구조화가 어렵지만, 다음의 2계층 구조(hybrid storage system)를 사용함으로써 문제를 해결하였다.[10] 첫 번째 계층은 원본 데이터(raw video/signal data)를 object storage (S3-compatible)에 저장하고, 파일 무결성 보존을 위해 SHA-256 해시값을 별도로 저장하며, 연구자 접근은 secure URL 발급을 통한 비직접 접근 방식으로 관리한다. 두 번째 계층은 구조화된 메타데이터 테이블을 구축하는 것으로, videonystagmography (VNG)에서는 saccade latency, gain, slow-phase velocity, nystagmus direction 등을, video head impulse testing (vHIT)에서는 gain, asymmetry, catch-up saccade 여부 등을, cervical vestibular evoked myogenic potential (VEMP)에서는 P13/N23 latency와 amplitude ratio를, subjective visual vertical에서는 deviation angle과 trial variability를 주요 파라미터로 정의하였다. 이는 원본 데이터는 object storage에 보관하면서, 원본 데이터로부터 추출하여 정형화 가능한 파라미터(예: canal paresis, catch-up saccade 여부, gain

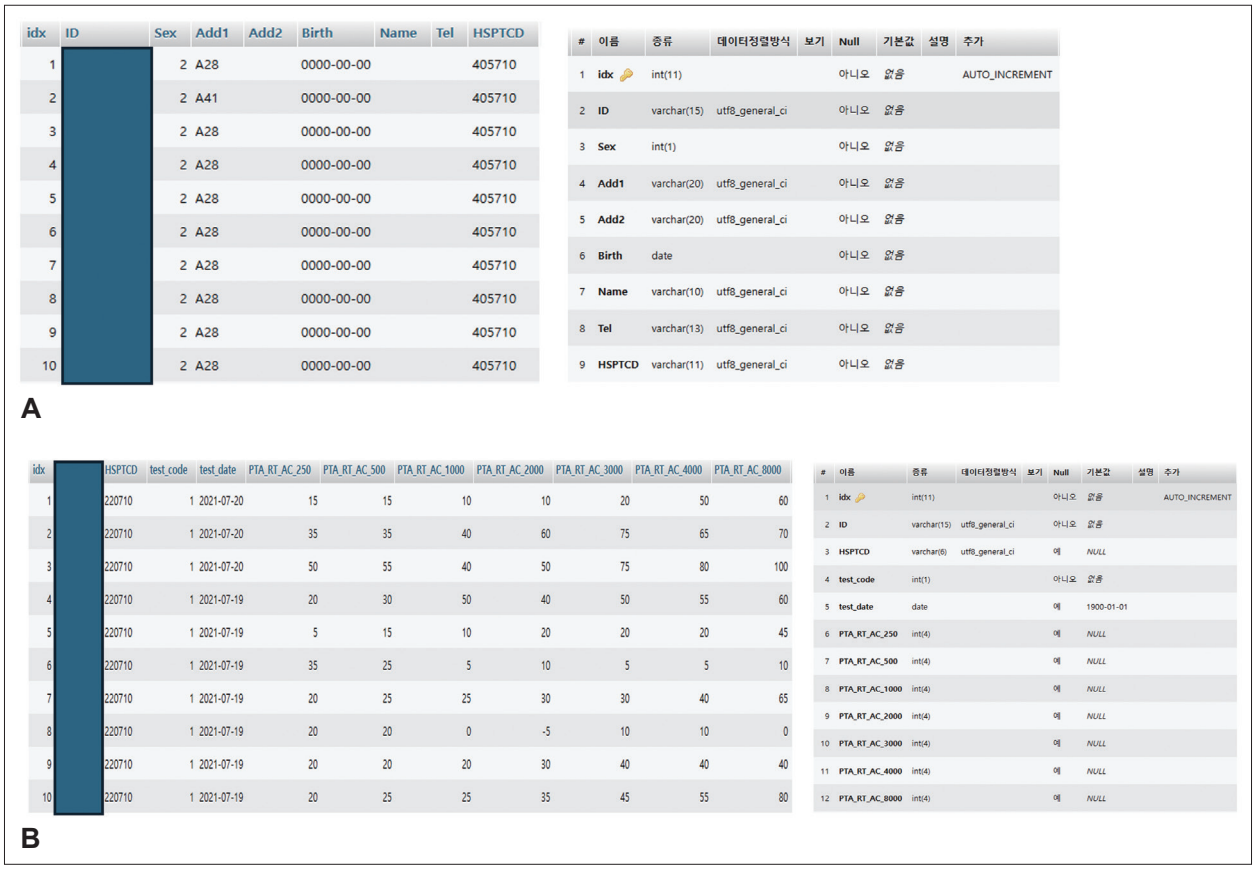

Fig. 2. Relational Database and Data Schema for Hearing Big Data. This figure illustrates the relational database and data schema designed for the Hearing Big Data Center Although the structure is simple, it provides significant extensibility. This relational database design offers advantages in terms of extensibility and integration when additional clinical information (e.g., vestibular function test data, general disease data such as blood tests) is added for patients included in the database. A: Profile table. B: Pure tone audiometry table.

값 등)를 메타테이블에 연동하는 하이브리드 구조이다. 이러한 구조는 데이터 분석 또는 AI 모델 학습에서 비정형 데이터는 필요할 때만 호출하고, 정형화된 핵심 파라미터는 SQL 기반 통계분석 및 AI feature engineering에 사용할 수 있게 하는 효율적인 형태이다.[11]

데이터베이스 조회 성능 향상을 위해 patient_id, visit_date, test_type, frequency (PTA) 등 조회 빈도가 높은 속성에 B-tree 인덱스를 적용하고, 외래키(FK)에도 보조 인덱스를 생성하여 조인 성능을 최적화하였다. PTA long-form 구조는 pivot 연산 없이 WHERE 조건 기반 조회가 가능하도록 설계하였으며, 반복 계산을 줄이기 위해 materialized view를 운영하였다. VNG 메타데이터처럼 대용량 비정형 요소가 포함된 경우에는 필요한 부분만 불러오는 lazy-loading 전략을 사용하여 효율성을 높였다. 백업 전략은 매일 전체 백업과 시간 단위 증분 백업을 병행하고, write-ahead logging 보관을 통해 시점 복구(point-in-time recovery)를 지원하였다. 또한 주요 검사값에는 domain constraint (예: PTA -10 – 120 dB, vHIT gain 0-2.0)를 적용하여 무결성을 확보하고, 위반 시 자동 알림이 발생하도록 설정하였다.

데이터 구축 과정에서는 다양한 현실적인 문제들이 발생하였다. 같은 환자가 동일 날짜에 여러 검사를 시행하는 경우 중복 저장 가능성이 있어, patient_id + visit_id + test_type + timestamp 조합을 이용한 unique key 규칙을 도입하여 중복을 제거하였다. 병원별 검사 포맷 차이로 인해 용어, 이미지 품질, 좌우 표기 방식 등이 달라 불일치가 빈번했으며, 이를 해결하기 위해 템플릿 분류기 기반 optical character recognition 전처리와 threshold rule engine을 적용하였다. 누락값 및 이상값과 같은 품질 문제에 대해서는 missing pattern 분석과 interquartile range·Z-score 기반 이상치 탐지 규칙을 도입하여 자동 보정 및 검증 체계를 마련하였다. 또한 비정형 파일의 저장 경로와 메타데이터 이름 규칙이 일관되지 않아 연동 오류가 발생했기 때문에, ingestion 단계에서 naming rule을 자동으로 적용하고 해시값 기반 파일-메타데이터 매칭 체계를 구축하였다.[8]

보건의료데이터베이스 구축의 일반적 사항

이비인후과 영역의 데이터를 포함한 의료 데이터는 그 데이터가 민감한 개인정보임을 반드시 기억하여야 한다. 우리나라에서는 2020년 9월에 보건복지부와 개인정보보호위원회에서 보건의료데이터 활용 가이드라인을 제정하고 꾸준히 개정해 왔으며 2024년 12월에 마지막으로 업데이트 되었다.[12] 연구자 입장에서 주의하여야 할 점은 개인정보 보호법에 위배되지 않도록 보건의료데이터를 활용하는 것이며, 이때 가명정보와 익명정보를 구분하여 사용해야 한다.

가명정보는 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 암호화 방법으로 개인을 직접적으로 식별할 수 없도록 해 놓은 정보를 말한다. 가명정보는 추가정보(예: 암호화 키)와 결합하면 개인을 다시 식별할 수 있는 가능성이 있다. 이 때 암호화 키와 같은 추가적 정보없이 특정 개인을 알아볼 수 없도록 처리하는 것을 '가명정보 처리'라고 하며, 보건정보빅데이터 가이드라인에서는 통계작성, 과학적 연구, 공익적 기록보존에 한정하여 정보주체의 동의 없이 가명정보를 생성, 이용, 제공 등 처리하거나 결합할 수 있다고 명시한다.[12]

익명정보란 시간, 비용, 기술 등을 합리적으로 고려할 때 다른 정보를 사용하더라도 더 이상 개인을 알아볼 수 없는 정보를 말하며, 특정 개인을 식별할 수 있는 모든 정보를 완전히 제거하거나 변환하여 어떠한 방법으로도 개인을 식별할 수 없다는 점에서 가명정보와 구분된다. 그러나 익명정보는 개인식별이 불가능해야 한다는 이유로 나이, 성별 등의 인구학적 정보의 손실이 많기 때문에 의료데이터를 분석해야 하는 관점에서 분석 결과의 임상적 유용성에 제한이 있다는 한계가 있다.

보건의료데이터베이스의 보안, 접근제어 및 로그 관리

의료 데이터베이스를 구축하고 운영하는 과정에서 보안과 프라이버시 보호는 가장 중요한 요소 중 하나이며, 저자들은 이를 위해 다층적 접근을 적용하였다. 첫째, 역할 기반 접근 제어(role-based access control)와 최소 권한 원칙을 엄격히 적용하여 데이터 접근 권한을 세분화하였다. 연구책임자, 전담 분석자, 시스템 관리자 등 역할별로 권한을 구분하여, 예를 들어 시스템관리자는 데이터 수정 및 관리 권한은 부여하되 분석 산출물에는 접근할 수 없도록 하고, 분석자는 raw data에는 접근하지 못하도록 설계하였다. 모든 접근 행위는 SQL 기반 자동 로그로 기록되며, 월 1회 정기적인 보안점검을 통해 이상 여부를 확인하도록 하였다.

둘째, 네트워크 보안 측면에서 청각빅데이터센터는 실제 개인정보가 포함된 데이터 분석을 위해서는 인터넷 완전 차단망(air-gapped environment)에서 운영되어 외부 네트워크와의 연결을 원천적으로 차단하였다. 데이터 반출이 필요할 경우 자동 diff-check 시스템을 통해 변경 내역을 검증하고, 보안 담당자 2인의 승인을 받아야 반출이 가능하도록 이중 승인 절차를 적용하였다. 또한 모든 외부 저장장치(USB, 휴대용 디스크 등)의 사용을 원칙적으로 제한하였다.

셋째, 데이터 전송 및 저장 과정에서 강력한 암호화 체계를 도입하였다. 저장 시에는 advanced encryption standard-256 기반 암호화를 적용하고, 데이터 전달 시에는 transport layer security 1.3 보안 프로토콜을 활용하여 전송 경로에서의 위험을 최소화하였다. 영상 및 음성 데이터를 포함한 비정형 데이터는 object storage에 보관하되, bucket-level 정책을 적용하여 접근 권한을 세밀하게 관리하였으며, presigned URL은 매우 짧은 유효기간을 부여하여 일시적 접근만 허용하였다.

넷째, 가명정보의 안전성을 평가하기 위해 재식별 위험 분석(re-identification risk assessment) 절차를 운영하였다. 본 연구에서는 k-anonymity 기준($k \geq 5$)을 충족하는지 여부를 우선 확인하고, quasi-identifier (나이, 지역, 방문 날짜 등)의 조합이 식별 위험을 높이지 않는지 분석하였다. 위험이 높다고 판단될 경우 날짜 정보를 연도 단위로 변환하거나, 나이를 범주화(age banding)하는 등 추가적인 가명처리를 수행하여 안전성을 보완하였다.

마지막으로, 감시(audit) 및 이상 탐지 체계를 구축하여 운영 중 발생할 수 있는 보안 위협을 실시간 관리하였다. 반복적이거나 비정상적인 조회 패턴(예: 특정 환자의 정보만 지속적으로 조회하는 경우)을 자동 탐지하는 규칙 기반 감시체계를 운영하였으며, 연구 단계에서는 머신러닝 기반 접근 이상 탐지 모델을 시험 적용하여 잠재적 보안 리스크를 조기에 발견할 수 있는 가능성을 확인하였다. 이와 같은 다층적 보안 체계는 의료 데이터베이스의 민감성을 고려할 때 필요한 수준의 보호 조치를 제공하며, 데이터 활용과 보안의 균형을 유지한 운영 모델로 기능하였다.

가명화 처리의 주의사항 및 가명정보의 결합

임상 연구자가 의료정보 데이터베이스의 구축과 분석 시 주의해야 할 점 중 하나는 가명처리를 하는 과정이다. 개인정보 처리자는 통계작성, 과학적 연구, 공익적 기록보존에 한정

하여 정보주체의 동의 없이 가명정보를 생성, 이용, 제공 등 처리하거나 결합할 수 있다고 되어 있으나 가명화 처리는 민감한 보건의료 데이터를 안전하게 활용하기 위해 필수적인 과정으로, 이를 효과적으로 수행하기 위해 몇 가지 주요 원칙과 절차를 준수해야 한다. 첫째, 가명화 과정은 명확한 목적 설정에서 시작된다. 데이터 활용 목적은 통계작성, 과학적 연구, 공익적 기록보존 등으로 제한되며, 각 목적에 부합하는 데이터 처리 방식을 사전에 설계해야 한다. 특히, 데이터 처리에 앞서 가명화 대상 데이터의 식별 위험성을 검토하고, 처리 환경의 특성을 고려하여 적절한 보호조치를 적용해야 한다. 예를 들어, 의료 영상 데이터의 경우 특정 개인의 신체적 특징을 마스킹하거나 식별 가능한 메타데이터를 삭제함으로써 재식별 위험을 줄일 수 있다. 이러한 과정에서 생성된 추가정보(예: 암호화 키 등)는 원본 데이터와 분리하여 저장하고 필요시 이를 즉시 파기해야 한다.

가명정보 결합은 보건의료 데이터를 융합하여 과학적 연구 및 공공 목적을 달성하는 데 중요한 도구로 사용될 수 있다. 그러나 데이터 결합 과정에서 식별 위험성이 증가할 가능성이 있으므로 결합 환경에 대한 엄격한 통제가 필수적이다. 모든 데이터 결합 작업은 폐쇄형 분석 환경에서 이루어져야 하며, 결합 후 생성된 데이터는 추가적인 가명화 과정을 통해 식별 가능성을 최소화해야 한다. 결합 작업에 사용되는 결합키는 데이터의 연계성을 유지하면서도 개인을 식별할 수 없도록 설계되어야 하며, 이러한 결합키는 관리 전문기관을 통해 엄격히 관리된다. 특히 반복 결합이 필요한 경우, 각 결합 단계마다 데이터의 안전성을 재검토하여 식별 위험을 줄이는 데 중점을 두어야 한다. 이러한 절차는 데이터 활용의 효율성을 유지하면서도 정보주체의 권리를 보호하는 데 기여한다.

결론적으로, 가명화와 데이터 결합은 현대 의료 데이터 활용의 핵심 기술로 자리 잡았지만, 이를 성공적으로 수행하기 위해서는 철저한 계획과 지속적인 관리가 필요하다. 데이터의 민감성 및 활용 목적에 따른 맞춤형 가명화 처리를 적용하고, 결합 과정에서 발생할 수 있는 식별 위험성을 사전에 평가함으로써 데이터의 안전성을 확보할 수 있다. 이를 통해 데이터 활용의 잠재성을 극대화하면서도 개인정보 보호라는 윤리적 책임을 준수할 수 있을 것이다.

AI 모델적용의 예 및 확장성

구축된 데이터베이스는 정형·비정형 정보를 모두 포함하고 있어 다양한 형태의 인공지능 모델 개발에 활용될 수 있다. 먼저 정형데이터 기반 모델 측면에서 PTA long-form 구조는 threshold 추정이나 청력 회복 예후 예측 모델에서 feature engineering이 용이하며, VNG 메타데이터는 Random Forest나 XGBoost 등에서 중요한 변수로 반복적으로 활용된다. 또한 방문 단위로 정리된 구조는 시계열 특성을 반영할 수 있어 long short-term memory (LSTM)이나 Transformer 기반 모델의 입력 형태로 자연스럽게 연결된다.

비정형 데이터도 AI 적용 가능성이 크다. VNG 동영상은 안진 패턴을 탐지하는 모델 개발에 활용될 수 있으며, vHIT raw trace는 CNN-LSTM 모델을 이용한 catch-up saccade 검출에 적용 가능하다. 환자의 증상 설명 음성 파일은 speech-to-text 변환 후 natural language processing embedding으로 처리하여 정량적 특징을 추출할 수 있다. 또한 본 데이터베이스는 다중모달(multimodal) AI 모델 개발에도 적합하다. 정형 검사 결과, 영상·신호 데이터, 임상 텍스트를 시간순으로 통합한 JavaScript Object Notation 기반 timeline 구조를 구성하면, 이를 입력으로 활용하는 multimodal transformer 모델을 구현할 수 있으며 진단 및 예후 예측 성능 향상에 기여할 수 있다.[13] 이러한 멀티모달 정보 융합의 중요성은 최근 스마트 헬스케어 분야에서 점차 강조되고 있다.[13]

AI 모델 개발을 위해 Extract (추출), Transform (변환), Load (적재) 파이프라인을 자동화하여 정형 데이터는 SQL 기반으로 추출하고, 비정형 데이터는 presigned URL을 사용해 안전하게 호출하도록 구성하였다. 변환 단계에서는 결측값 처리, 정규화, windowing 등 기본 전처리를 수행하고, 임상 항목은 SNOMED·LOINC과 같은 표준 용어체계에 매핑하여 분석 일관성을 높였다. PTA, VNG, vHIT 등 주요 검사에서는 slope, asymmetry index, gain asymmetry 등 모델 학습에 유용한 feature를 자동 생성하는 feature engineering 모듈을 운영하였다. 이 과정을 통해 tabular 모델(XGBoost, CatBoost), 시계열 모델(LSTM·Transformer), 그리고 영상·텍스트·수치 정보를 결합한 다중모달 모델을 위한 학습 데이터셋을 손쉽게 구성할 수 있었다. 특히 영상 CNN과 tabular 모델의 결합 또는 BERT·LLM 기반 임상 텍스트 embedding의 수치형 feature 통합은 복잡한 임상 패턴을 반영하는 데 효과적이었다.[12]

한편, 데이터베이스 일부 요소는 공통데이터모델(CDM) 구조와의 매핑이 가능하도록 설계되어 있어, 향후 다기관 공동연구와 표준화된 분석 도구 적용에 유리하다. 특히 영상 및 검사 메타데이터를 OMOP-CDM 또는 Radiology-CDM 구조에 연계하는 시도는 국내외에서 활발히 이루어지고 있으며,[13] 청각 및 전정 데이터에서도 유사한 표준화 노력이 요구된다. 최근 청각 분야에서도 데이터 표준화와 상호운용성을 위한 국제 논의가 활발해지고 있어, 향후 청각빅데이터센터와 같은 구축 경험이 이러한 표준화 논의에 기여할 수 있

을 것으로 기대된다.[14,15]

마무리

바야흐로 인공지능의 시대이다. 인공지능 기술의 근간이 되는 데이터 과학에 대한 이해가 무엇보다 중요하며, 의료 분야에서 무궁무진하게 축적되고 있는 방대한 데이터를 데이터사이언스의 관점에서 바라보고 이를 적극적으로 활용하기 위한 지혜가 필요하다. 의료 데이터베이스 구축은 정형 및 비정형 데이터를 효율적으로 관리하고, 확장 가능성과 표준화를 고려한 체계적 설계를 바탕으로 이루어져야 한다. 특히 의료 데이터의 민감성을 고려하여 개인정보 보호와 가명화 처리를 철저히 준수하면서도, 데이터의 가치를 최대한 보존할 수 있는 적절한 활용 전략을 마련해야 한다. 궁극적으로 의료 데이터를 데이터사이언스적 접근법으로 현명하게 설계하고 구축하는 것이 의료 연구 및 진료의 발전을 이끄는 핵심이 될 것이다. 본 지가 이비인후과 연구자들이 본인의 임상 환경에 적합한 데이터베이스를 설계하고, 향후 인공지능 및 다중모달 분석으로 확장하는 과정에서 실질적인 "how I do it" 지침으로 활용되기를 기대한다.

Acknowledgments

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술 연구개발사업 지원에 의하여 이루어진 것임(과제고유번호: RS-2025-25460099).

Author Contribution

Conceptualization: Young Joon Seo, Tae Hoon Kong. Investigation: Young Joon Seo. Methodology: Tae Hoon Kong. Project administration: Young Joon Seo, Tae Hoon Kong. Supervision: Young Joon Seo. Writing—original draft: Tae Hoon Kong. Writing—review & editing: Tae Hoon Kong, Young Joon Seo.

ORCIDs

Tae Hoon Kong https://orcid.org/0000-0002-5612-5705
Young Joon Seo https://orcid.org/0000-0002-2839-4676

REFERENCES

1) Schwab K. The fourth industrial revolution. Geneva: World Economic Forum;2016.
2) World Economic Forum. Technological innovation [online] [cited 2025 Feb 26]. Available from: URL: https://www.weforum.org/focus/fourth-industrial-revolution.
3) Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529(7587):484-9.
4) Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017;550(7676):354-9.
5) Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349(6245):255-60.
6) Demchenko Y, Grosso P, De Laat C, Membrey P. Addressing big data issues in scientific data infrastructure. Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS); 2013 May 20-24; San Diego, CA, USA. Piscataway: IEEE;2013. p.48-55.
7) Elmasri R, Navathe SB. Fundamentals of database systems. 7th ed. Boston: Pearson;2016.
8) Shortliffe EH, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine. 4th ed. London: Springer;2014.
9) Codd EF. A relational model of data for large shared data banks. Commun ACM 1970;13(6):377-87.
10) Samra HE, Li A, Soh B. Design of a clinical database to support research purposes: challenges and solutions. Int J Adv Appl Sci 2021;8(3):21-9.
11) Darmont J, Olivier E. A complex data warehouse for personalized, anticipative medicine. Proceedings of the 17th International Conference of the Information Resources Management Association (IRMA 2006); 2006 May 21-24; Washington, DC: USA. Hershey: Information Resources Management Association;2006. p.685-7.
12) Ministry of Health and Welfare, Personal Information Protection Commission [Guidelines for the use of healthcare data]. Sejong, Seoul: Ministry of Health and Welfare, Personal Information Protection Commission;2024. Korean
13) Shaik T, Tao X, Li L, Xie H, Velásquez JD. A survey of multimodal information fusion for smart healthcare: mapping the journey from data to wisdom. Inf Fusion 2024;102:102040.
14) Park C, You SC, Jeon H, Jeong CW, Choi JW, Park RW. Development and validation of the radiology common data model (R-CDM) for the international standardization of medical imaging data. Yonsei Med J 2022;63(Suppl):S74-83.
15) Vercammen C, Heinrich A, Lesimple C, Paglialonga A, Wasmann JA, Buhl M. Data standards in audiology: a mixed-methods exploration of community perspectives and implementation considerations. Int J Audiol. In press 2026.