



Design and Development of Medical Databases

Tae Hoon Kong^{1,2,3,4} and Young Joon Seo^{1,2,3,4}

Departments of ¹Otorhinolaryngology-Head and Neck Surgery, ²Medical Informatics and Biostatistics, and ³Bio-AI Convergence, Yonsei University Wonju College of Medicine, Wonju; and

⁴Research 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 ¹이비인후과학교실, ²의료정보통계학과, ³바이오-AI 융합학과, ⁴청각재활연구소

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

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 2026;69(4):198-203

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

Introduction

Since the Fourth Industrial Revolution was declared at the World Economic Forum in 2016, rapid technological innovation has become evident in the six core domains of big data, artificial intelligence (AI), robotics, the Internet of Things,

unmanned aerial systems, three-dimensional printing, and nanotechnology.^{1,2)} In 2016, Google DeepMind demonstrated the capabilities of AI by defeating Lee Sedol, one of the world's top Go players, with AlphaGo, highlighting AI performance in a domain traditionally regarded as requiring human intuition and creativity.³⁾ Tesla has similarly been recognized not only for advancing electric vehicles but also for pioneering AI-driven autonomous driving technologies that have transformed the automotive industry.⁴⁾

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the widespread introduction of generative AI, including large language models such as ChatGPT, AI tools are now routinely used to enhance productivity, functioning in many cases as virtual personal assistants.⁴⁾ Although the evolution of AI has been driven by advances in computing technologies, algorithms, and platforms, its most important underlying factor is the development of technologies for large-scale data collection and management.⁵⁾ Improvements in computing power and cloud-based infrastructures have enabled not only efficient storage and retrieval of vast datasets but also real-time data analysis and visualization.⁵⁾

These trends are often summarized by the “5V” characteristics of big data—Variety, Volume, Velocity, Value, and Veracity—reflecting the intention to collect and utilize heterogeneous data generated across diverse domains, including text, logs, and images, rather than focusing solely on large numerical datasets.⁶⁾

In medicine, particularly in otolaryngology, clinical data frequently take unstructured forms, such as hearing test results, vestibular function test outputs, radiologic and magnetic resonance imaging, and electronic medical records (EMRs). As a result, data structures and formats vary widely depending on institutions, devices, and testing protocols, leading to limited reusability from the perspectives of long-term research and AI development. Against this background, the authors have accumulated practical expertise through projects such as the Hearing Big Data Center of the Korean Audiological Society, Common Data Model (CDM) implementation initiatives, and the construction of vestibular function test databases related to dizziness. Based on these experiences, this article introduces design principles and concrete implementation examples for otolaryngology databases capable of simultaneously managing structured and unstructured data.

Understanding Databases: Structured and Unstructured Data

A database is conventionally defined as a collection of values represented as quantitative or qualitative variables.⁷⁾ In the era of big data, this definition has expanded to encompass data types that are difficult to classify strictly as quantitative or qualitative, including text, audio, images, and video. Nevertheless, meaningful data analysis requires that such information ultimately be organized into analyzable forms, underscoring the importance of understanding the distinction

between structured and unstructured data.⁷⁾

Structured data follow predefined data models or schemas and are organized within systematic frameworks, allowing efficient searching and analysis due to their quantitative characteristics. Such data are typically stored in data warehouses using structured query language (SQL)-based relational database systems.⁷⁾ In contrast, unstructured data refer to all other data types that do not conform to predefined schemas and may exist as text, audio, images, or video. In otolaryngology and medicine more broadly, most diagnostic reports, imaging studies, and EMR-based clinical narratives fall into this category.⁸⁾ Although unstructured data offer flexibility, they are inherently difficult to search and analyze systematically due to the absence of explicit schemas.⁸⁾

Therefore, to integratively manage the diverse structured and unstructured data generated in otolaryngology, careful database design is required to determine how these two data types should be structured, linked, and utilized in a complementary manner.⁸⁾

Relational Database Architecture and Data Schemas

Since its initial proposal by Codd⁹⁾ in 1970, the relational database model has remained the most widely adopted database framework. In relational databases, data attributes of individual entities are represented in two-dimensional tables, and relationships between entities are defined using relational algebra and relational calculus. Fig. 1 illustrates an example of a relational database model in which “Student,” “Professor,” and “Course Offering” entities are represented as separate tables. Relationships between the “Student” and “Professor” entities are defined through a “Guidance” relationship using a shared “Professor ID” attribute, while course offerings and registrations are linked through departmental attributes. This structure demonstrates how relational databases organize entity attributes into concept-based tables with explicit relationships.

A data schema provides a comprehensive specification of database structure and constraints. It defines attributes, entities composed of attribute sets, relationships among entities, and integrity constraints that must be maintained.⁷⁾ In relational databases, the relational schema is an essential design element, as it explicitly defines how entities and their relationships are structured, ensuring data consistency, integrity, and extensibility.⁷⁾

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

A

Professor_code	Professor_Name	Department	Subject_No	Subject_Name	Department
1234	Kong	Statistics	AA1234	Math1	Statistics
4077	Seo	Computer	AA5678	Medicine	Informatics
5456	Kim	Statistics	AA2345	English	Common
6668	Lee	Informatics	AB3456	Math2	Statistics
8723	Park	Computer	AC9876	Data1	Computer

B **C**

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.

Medical Database Construction: Otolaryngology Cohorts and Hybrid Architecture

Many researchers construct datasets to test specific research hypotheses; however, building databases solely for single hypotheses is inefficient.¹⁰⁾ Consequently, disease-specific prospective or retrospective cohorts are often established and reused as databases. Such cohort data are frequently stored as spreadsheet-based case report forms, which limits scalability and long-term usability from a database design perspective.¹⁰⁾

During the construction of the Hearing Big Data Center in 2020, the authors designed a relational database and data schema with a simple yet highly extensible structure (Fig. 2). This database integrated approximately 120000 pure-tone audiometry (PTA) records, over 8000 speech audiometry tests (SRT/WRS), tens of thousands of diagnostic codes, and associated metadata. Because audiometric test sheets varied widely across institutions and were largely unstructured, direct storage of raw formats would have hindered SQL-based analysis. To address this, thresholds were manually digitized, patient and test tables were linked in a one-to-many structure, and frequency-specific thresholds were stored in a long-form format. This design supports future extensions—such as additional frequencies, bone and air conduction data, and masking information—while maximizing analytical efficiency

through SQL-based queries.

For vestibular function test data, which inherently include large volumes of unstructured data, a two-tier hybrid storage system was adopted.¹⁰⁾ The first tier stores raw video and signal data in S3-compatible object storage, with SHA-256 hash values preserved to ensure file integrity and access controlled through secure URLs. The second tier consists of structured metadata tables defining clinically relevant parameters, such as saccade latency, gain, slow-phase velocity, and nystagmus direction for videonystagmography (VNG); gain, asymmetry, and catch-up saccades for video head impulse testing (vHIT); P13/N23 latency and amplitude ratios for cervical vestibular evoked myogenic potential (VEMP); and deviation angles and trial variability for subjective visual vertical testing. This hybrid structure allows unstructured data to be accessed only when necessary, while structured parameters are readily available for SQL-based statistical analyses and AI feature engineering.¹¹⁾

To optimize query performance, B-tree indices were applied to frequently queried attributes such as patient identifiers, visit dates, test types, and PTA frequencies, with auxiliary indices created for foreign keys to improve join efficiency. Materialized views were employed to reduce repetitive computations, and lazy-loading strategies were used for large unstructured components. Backup strategies combined daily full backups with hourly incremental backups, supporting point-in-time recovery through write-ahead log retention. Domain

idx	ID	Sex	Add1	Add2	Birth	Name	Tel	HSPICD
1		2	A28		0000-00-00			405710
2		2	A41		0000-00-00			405710
3		2	A28		0000-00-00			405710
4		2	A28		0000-00-00			405710
5		2	A28		0000-00-00			405710
6		2	A28		0000-00-00			405710
7		2	A28		0000-00-00			405710
8		2	A28		0000-00-00			405710
9		2	A28		0000-00-00			405710
10		2	A28		0000-00-00			405710

#	이름	종류	데이터형식	보기	Null	기본값	설명	추가
1	idx	int(11)			아니오	없음		AUTO_INCREMENT
2	ID	varchar(15)	utf8_general_ci		아니오	없음		
3	Sex	int(1)			아니오	없음		
4	Add1	varchar(20)	utf8_general_ci		아니오	없음		
5	Add2	varchar(20)	utf8_general_ci		아니오	없음		
6	Birth	date			아니오	없음		
7	Name	varchar(10)	utf8_general_ci		아니오	없음		
8	Tel	varchar(13)	utf8_general_ci		아니오	없음		
9	HSPICD	varchar(11)	utf8_general_ci		아니오	없음		

idx	HSPICD	test_code	test_date	PTA_RT_AC_250	PTA_RT_AC_500	PTA_RT_AC_1000	PTA_RT_AC_2000	PTA_RT_AC_3000	PTA_RT_AC_4000	PTA_RT_AC_8000
1	220710	1	2021-07-20	15	15	10	10	20	50	60
2	220710	1	2021-07-20	35	35	40	60	75	65	70
3	220710	1	2021-07-20	50	55	40	50	75	80	100
4	220710	1	2021-07-19	20	30	50	40	50	55	60
5	220710	1	2021-07-19	5	15	10	20	20	20	45
6	220710	1	2021-07-19	35	25	5	10	5	5	10
7	220710	1	2021-07-19	20	25	25	30	30	40	65
8	220710	1	2021-07-19	20	20	0	-5	10	10	0
9	220710	1	2021-07-19	20	20	20	30	40	40	40
10	220710	1	2021-07-19	20	25	25	35	45	55	80

#	이름	종류	데이터형식	보기	Null	기본값	설명	추가
1	idx	int(11)			아니오	없음		AUTO_INCREMENT
2	ID	varchar(15)	utf8_general_ci		아니오	없음		
3	HSPICD	varchar(8)	utf8_general_ci		예	NULL		
4	test_code	int(1)			아니오	없음		
5	test_date	date			예	1900-01-01		
6	PTA_RT_AC_250	int(4)			예	NULL		
7	PTA_RT_AC_500	int(4)			예	NULL		
8	PTA_RT_AC_1000	int(4)			예	NULL		
9	PTA_RT_AC_2000	int(4)			예	NULL		
10	PTA_RT_AC_3000	int(4)			예	NULL		
11	PTA_RT_AC_4000	int(4)			예	NULL		
12	PTA_RT_AC_8000	int(4)			예	NULL		

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.

constraints were applied to key parameters to ensure data integrity, with automated alerts triggered upon violations.

General Considerations for Healthcare Database Construction

Healthcare data, including those generated in otolaryngology, are inherently sensitive personal information and must be handled accordingly. In South Korea, healthcare data utilization guidelines were first established in September 2020 by the Ministry of Health and Welfare and the Personal Information Protection Commission and have been continuously revised, with the most recent update issued in December 2024.¹²⁾ Researchers must ensure that healthcare data are used in compliance with privacy laws and must distinguish between pseudonymized and anonymized data. Pseudonymized data are processed through techniques such as deletion or replacement of identifying information to prevent direct identification, although re-identification remains possible

when combined with additional information.¹²⁾

Anonymized data, by contrast, are processed so that individuals cannot be identified by any reasonably available means; however, this often results in substantial loss of demographic information, limiting clinical utility.¹²⁾

Security, Access Control, and Audit Logging in Healthcare Databases

Security and privacy protection are paramount in medical database construction and operation. The authors implemented a multilayered security approach. First, role-based access control and the principle of least privilege were strictly applied, differentiating permissions among principal investigators, analysts, and system administrators. All access activities were recorded through automated SQL-based logging and reviewed during monthly security audits.

Second, the Hearing Big Data Center was operated within an air-gapped environment to prevent external network ac-

cess when analyzing identifiable data. Data export required automated difference checks and dual authorization by security officers. Third, strong encryption mechanisms were applied during data storage (AES-256) and transmission (TLS 1.3). Unstructured data stored in object storage were governed by bucket-level access policies and accessed only through short-lived presigned URLs.

Fourth, re-identification risk assessments were conducted using k -anonymity criteria ($k \geq 5$), with additional pseudonymization measures applied when risks were identified. Finally, audit and anomaly detection systems were implemented to monitor abnormal access patterns, including exploratory application of machine learning-based intrusion detection models.

Considerations for Pseudonymization and Data Combination

Pseudonymization is a critical process in medical database construction and analysis. Although pseudonymized data may be processed without individual consent for statistical analysis, scientific research, and public interest purposes, strict procedural safeguards are required.¹²⁾ Pseudonymization begins with clear definition of data use purposes, followed by assessment of identification risks and application of appropriate technical and administrative protections. Additional information, such as encryption keys, must be stored separately and destroyed when no longer required.

Data combination can enhance scientific and public value but increases identification risk. Therefore, all data linkage must be performed in controlled environments, with additional pseudonymization applied to combined datasets. Linkage keys must preserve relational integrity while preventing individual identification, and repeated linkage requires reassessment of identification risks at each stage. These procedures balance efficient data utilization with the protection of data subjects' rights.

Examples and Scalability of AI Model Applications

The constructed database integrates both structured and unstructured information, enabling diverse AI applications. Structured data support feature engineering for predictive models of hearing thresholds and prognosis, while visit-level organization facilitates time-series modeling using recurrent

or transformer-based architectures. Unstructured data, such as VNG videos and vHIT raw traces, can be used to develop deep learning models for nystagmus pattern recognition and catch-up saccade detection. Speech recordings of symptom descriptions may be converted into text and processed using natural language processing embeddings.

Moreover, the database is well suited for multimodal AI models. By organizing structured test results, imaging and signal data, and clinical text into time-ordered JavaScript Object Notation-based timelines, multimodal transformer models can be implemented to improve diagnostic and prognostic performance.¹³⁾ Automated Extract, Transform, Load pipelines were developed to extract structured data via SQL and access unstructured data securely using presigned URLs, with standardized terminology mapping to systems such as SNOMED and LOINC to ensure analytical consistency.

Finally, portions of the database were designed to be compatible with CDM frameworks, facilitating future multicenter collaborations and standardized analyses. Ongoing international efforts toward data standardization in audiology and vestibular research highlight the importance of such interoperability.^{14,15)}

Conclusion

In the era of AI, understanding data science and database design is fundamental. Medical databases must be constructed using systematic designs that efficiently manage structured and unstructured data while ensuring scalability and standardization. At the same time, strict adherence to privacy protection and pseudonymization principles is essential to preserve the value of medical data without compromising ethical responsibilities. Ultimately, thoughtful database design from a data science perspective will play a central role in advancing medical research and clinical practice. This article is intended to serve as a practical “how I do it” guide for otolaryngology researchers seeking to design databases tailored to their clinical environments and to expand them toward AI-driven and multimodal analyses.

Supplementary Materials

Korean translation of this article is available with the Online-only Data Supplement at <https://doi.org/10.3342/kjorl-hns.2025.00122>.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Wel-

fare, Republic of Korea (grant number: 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

- Schwab K. The fourth industrial revolution. Geneva: World Economic Forum;2016.
- World Economic Forum. Technological innovation [online] [cited 2025 Feb 26]. Available from: URL: <https://www.weforum.org/focus/fourth-industrial-revolution>.
- 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.
- 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.
- Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;349(6245):255-60.
- 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.
- Elmasri R, Navathe SB. Fundamentals of database systems. 7th ed. Boston: Pearson;2016.
- Shortliffe EH, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine. 4th ed. London: Springer;2014.
- Codd EF. A relational model of data for large shared data banks. *Commun ACM* 1970;13(6):377-87.
- 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.
- 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.
- 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
- 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.
- 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.
- 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.