KMBIG: Safeguarding Data Sharing with Advanced Anonymization and Risk Management

1st Icktae Kim
KM Data Division
Korea Institute of Oriental Medicine
Daejeon, Republic of Korea
kimit@kiom.re.kr

2nd Taehong Kim
KM Data Division
Korea Institute of Oriental Medicine
Daejeon, Republic of Korea
thkim@kiom.re.kr

2024 IEEE International Conference on Big Data and Smart Computing (BigComp) | 979-8-3503-7002-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIGCOMP60711.2024.00097

***Abstract*— In August 2020, South Korea's data activation legislation was updated, introducing "Guidelines for the Use of Health and Medical Data," thereby extending its reach to the medical sector. These guidelines delineate crucial protocols for individual medical data use, emphasizing anonymization and mandatory Data Review Committee oversight. Our research proposes a methodology within the KMBIG system for health data utilization, compliant with these stipulations. Through comprehensive analysis, we have developed specific anonymization techniques for various data types, enabling direct validation and swift identifiability risk assessment. The KMBIG system's deployment is anticipated to significantly enhance healthcare data efficacy, aligning with the regulatory framework**

Keywords—Data Sharing and Utilization, Personal Data Protection Act, Pseudonymization Dynamic Processing Module, De-identification Process, Privacy Data Risk Index

I. INTRODUCTION

Data is a fundamental asset in driving the Fourth Industrial Revolution, pivotal for economic enhancement and societal advancement. It fosters the creation of novel products, services, business strategies, public policies, and societal solutions. Yet, its utilization poses significant privacy risks, potentially infringing personal rights and safety. Global legislative efforts, like the European Union's General Data Protection Regulation (GDPR) and the United States California Consumer Privacy Act (CCPA), underscore a commitment to bolster personal data security.

In South Korea, the amended Personal Information Protection Act (2020) reflects a dual aim: safeguarding privacy and catalyzing data usage[1-3]. This legislation introduces key measures, including the institutionalization of pseudonymous and anonymous data for safer use, enhanced security protocols for data sharing, and expanded legal accountability for privacy violations[4]. Table 1 depicts the multifaceted approach of various Korean government agencies in data governance, illustrating a commitment to comprehensive data protection and management.

By 2022, specific guidelines for medical data usage were released, emphasizing the creation of data review committees and stringent criteria for pseudonymous information processing. Despite these robust frameworks, practical challenges in medical data sharing persist, ranging from the specifics of pseudonymization to the intricacies of review and management processes[5-8].

In practice, however, the application of these guidelines to medical data sharing is fraught with several challenges. Firstly, there is the critical issue of determining which personal information should be pseudonymized and establishing a method for doing so effectively. This involves not only identifying sensitive data but also deciding on the level and technique of pseudonymization that balances data utility with privacy. Secondly, the stability of de-identification post-pseudonymization must be evaluated to ensure that the risk of re-identification is minimal. This requires a set of standards and metrics to assess the strength of the pseudonymization applied and continual monitoring to maintain data privacy. Designing the data review committee's review process is another significant challenge. The process needs to be comprehensive, transparent, and efficient, ensuring that all pseudonymized data is rigorously evaluated for safety, ethical use, and compliance with legal standards. The committee must navigate complex decisions, often weighing the benefits of data sharing against the risks of potential privacy breaches. Furthermore, there are important considerations for data that does not pass the review and for data that is approved. For data that fails to meet the required standards, clear protocols must be established for its re-evaluation, modification, or disposal. Conversely, for data that passes the review, there must be guidelines for its subsequent use, sharing, and long-term management to ensure that it remains secure and its use remains ethical and legal. Lastly, the management of data after it has been shared is pivotal. Ensuring the integrity and confidentiality of data, monitoring its usage, and protecting it from unauthorized access or breaches are all part of the post-sharing data management challenge. This involves not only technical solutions but also policy and governance measures to oversee the lifecycle of shared data.

To address these challenges, the Korea Institute of Oriental Medicine has developed a system consisting of data anonymization, data visualization, data review support, and data

2375-9356/24/$31.00 ©2024 IEEE
DOI 10.1109/BigComp60711.2024.00097

management modules [9]. The main functions of the system are as follows:

TABLE 1. PERSONAL INFORMATION DATA GOVERNANCE BY KOREAN GOVERNMENT AGENCIES

	Data-Driven Governance Activation Act	Private Data	Personal Information Protection Act	Credit Information Act
Responsible Ministry	Ministry of the Interior and Safety	Ministry of Science, ICT and Future Planning	Personal Information Protection Commission	Financial Services Commission
Legal Basis	Act on the Activation of Data-Driven Administration	Basic Act on Intelligent Information Society	Personal Information Protection Act	Law on the Use and Protection of Credit Information (Sep, 20)
Objectives and Main Contents	Establishes a scientific administrative system and fosters collaborative data sharing among public agencies.	Focuses on data production, collection, distribution, utilization, standardization, and quality improvement.	Aims to create a foundation for handlers to process and utilize personal information anonymously for statistical and research purposes.	Promotes the sound development of the credit information industry, efficient utilization, and systematic management of credit information, while ensuring privacy protection.

- Data anonymization module: Anonymizes data based on user-defined conditions and levels.
- Data visualization module: Visualizes anonymized data to make the anonymization process easier to understand.
- Data review support module: Generates materials necessary for the data review committee's review, such as anonymization plans and de-identification stability assessments.
- Data management module: Performs data management tasks such as tracking data usage and detecting potential re-identification risks after data sharing.

The system has been evaluated by experts in personal information protection and has been found to be effective in addressing the challenges of sharing medical data while protecting personal information.

II. KMBIG: DATA SAHRING SYSTEM

A. Data Integration Analysis in KMBIG System

In implementing the KMBIG system, we conducted an extensive analysis of the data utilized within the system. The shared and utilized data of the KMBIG service includes health checkup-centered clinical data, disease-centered clinical data, personal generated data, and multi-national oriental quantitative data. To ensure the safe handling of this data, particularly considering the risk of personal information, we formed a group comprising personal information protection experts from various fields along with a data analysis team. This group was tasked with reviewing the methods of sharing and utilizing the data to ensure compliance with personal information protection guidelines.

The focus of the analysis was on assessing the potential for data recombination and its associated risks. This entailed a thorough examination of all data types present in the database, including both structured and unstructured data, time series data, and file-based data, looking closely at types, column names, file names, and extensions. The intent was to identify and mitigate any risks pertaining to personal information, ensuring that the KMBIG system's data sharing and utilization are both secure and compliant with the highest standards of data privacy. This strategy, including the specific data types and considerations, is encapsulated in Table 2.

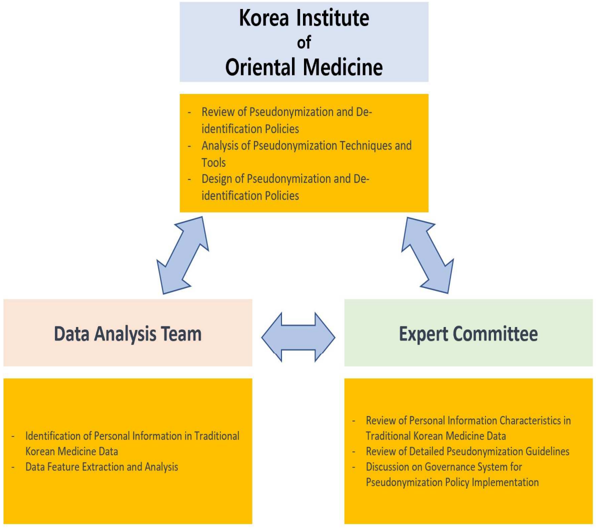

Korea Institute of Oriental Medicine
- Review of Pseudonymization and De-identification Policies
- Analysis of Pseudonymization Techniques and Tools
- Design of Pseudonymization and De-identification Policies
Data Analysis Team
- Identification of Personal Information in Traditional Korean Medicine Data
- Data Feature Extraction and Analysis
Expert Committee
- Review of Personal Information Characteristics in Traditional Korean Medicine Data
- Review of Detailed Pseudonymization Guidelines
- Discussion on Governance System for Pseudonymization Policy Implementation

Figure 1. Structure for Traditional Korean Medicine Data Analysis and Pseudonymization Review

TABLE 2. THE TYPE OF DATA IN KMBIG SYSTEM INCLUDE

Type	Data Description	Features
Hospital - Centered Medical Check-up Data	Medical Check-up Data from 5 Oriental Medicine Hospitals Combining Traditional and Western Medicine	Korean Traditional Medical Check-up Data Collection Program (Questionnaires, Devices), eCRF, Mobile APP Utilization
Disease-Centered Clinical Data	Disease-Specific Clinical Data	Disease-Specific Western/Traditional Medicine Clinical Key Indicator Measurement Standard Protocols
Personal Generated Health Data	Lifelog data, includes sleep, exercise, eating habits, and self-reported health information	Establishment of Individual-Generated Data Definition and Collection Protocol (Integration of Health Check-up EMR System and PHR Platform)
Multinational Traditional Medicine Quantitative Data	Sasang Constitution, Temperament and Character Inventory (TCI) Psychological Data	Development of International Collaborative Research SOP for Acquiring Korean-Chinese Constitutional/Psychological Clinical Data

TABLE 3. PSEUDONYMIZATION BY DATA TYPE IN KMBIG

Type	Pseudonymization Guideline by Data Type
Identifier	- Removal of identifiers is standard; however, substitution with serial numbers is permitted for healthcare data to ensure continuity in time-series and large-scale measurement data. - Enables research on the same individual's measurement data through substitution with serial numbers.
Critical Personal Information	- Addresses should be categorized at the town or neighborhood (eup, myeon, dong) level. - Age and birthdate information should be deleted. - Additional processing is required for data with a high potential for personal identification or considered as sensitive.
Data Value	- No measures needed for numerical measurement information. - No measures needed for medical professionals' observational input information. - No measures needed for health information generated by algorithms.
In Vitro Imaging Data	- Processing required for still images and videos showing physical appearance. - Complete removal or adequate mosaic processing or masking of all external physical features such as eyes, nose, mouth, tattoos, and other distinctive features. - Identification information such as patient numbers and names displayed in the images must be deleted or masked. - Identifiers in the DICOM headers and other metadata should be removed.
In Vivo Imaging Data	- Additional processing is required for 3D imaging information from CT scans. - Voice data usage is pending judgment and can be utilized based on individual consent.
Genomic Data	-Pending judgment with individual consent, except for: The presence or type of gene mutations related to widely known diseases. Removal of genetic mutation information of reproductive cells in neoplasms, except for new mutation information.
Non-Genomic Omics Data	- No additional processing is required.

The outcome of the analytical process led to the development of tailored anonymization guidelines for each type of data within the KMBIG system. This deliberate approach is intended to improve the effectiveness of anonymization techniques, enhancing the overall security and privacy of data.

De-identification protocols for healthcare data have been established based on varied data types, as systematically documented in Table 3. This initiative is a testament to the commitment to upholding data integrity and confidentiality while enabling ethical and compliant data utilization.

B. Enhancing Data Review Committee's Review Process and Stability Reporting

Under the "Guidelines for the Use of Medical Data" set forth by the Ministry of Health and Welfare along with the Personal Information Protection Commission, the data review committee is mandated to perform a comprehensive review to assure the safe and ethical application of pseudonymized medical data.

TABLE 4. DELIBERATION TOPICS FOR DATA REVIEW COMMITTEE [7]

Subject for Review	Description
The purpose of pseudonymization	Ensure the pseudonymization objective aligns with the original data collection and usage purpose. It must be clearly articulated and justified.
The adequacy of pseudonymization	Verify that the pseudonymization process sufficiently protects individual privacy and is designed to prevent re-identification, even when combined with other data.
Internal Utilization of Pseudonymized Data	Ascertain that the pseudonymized data is used within the institution in a safe, ethical manner, strictly for its intended purposes.
External Provision of Pseudonymized Data	Ensure that any provision of pseudonymized data to external entities is conducted ethically and safely, restricted to parties with a legitimate need and adequate privacy protections in place.

The committee, comprising at least 5 but no more than 15 members, primarily includes external experts, playing a pivotal role in safeguarding individual privacy and leveraging pseudonymized medical data for societal benefit. The review process unfolds as follows:

1) User Review Application: This initiates when a user intends to share data externally, necessitating verification of data suitability followed by a formal request submission to the committee.

2) Managerial Review Management: Upon receipt of a request, the committee manager assigns it for expert evaluation.

3) Expert Review Progress: Experts conduct a thorough assessment, considering the data sharing's purpose, de-identification adequacy, re-identification risks, and the overall potential benefits. Subsequently, they may approve, suggest modifications, or reject the request.

4) User Data Utilization Post-Approval: Approved data sharing enables users to proceed with data download and utilization for the agreed-upon purposes. Rejected applications may be resubmitted post-revisions.

We advocate for employing a stability report as a key tool in the review process. This report offers a detailed quantitative analysis of the risk associated with re-identifying individuals from the data set, thereby equipping experts with vital metrics to guide their review decisions.

The report uses three different algorithms to measure the risk: DCAP, CIO, and ROE[10]. DCAP is a statistical measure of the probability that a specific individual can be re-identified, given the data sharing request. CIO measures the overlap between the confidence intervals of the estimates of the true values of the parameters in the data. ROE measures the ratio of the estimates of the true values of the parameters in the data to the estimates of the values of the parameters in the data sharing request. The study also proposes the use of K-anonymity to assess the potential for re-identification. K-anonymity is a property of a dataset that indicates that each individual in the dataset is indistinguishable from at least K-1 other individuals[11-14]. A higher K-value indicates a lower risk of re-identification. The study's proposed process is designed to be efficient and effective in protecting the privacy of individuals while also facilitating the use of data for research and other purposes.

C. Pseudonymization Pipeline Module

The pseudonymization pipeline module is instrumental in appending pseudonymization elements to data or elevating the pseudonymization degree. When the user-applied pseudonymization is insufficient, it amplifies the risk of identification, necessitating a more robust pseudonymization process for each specific data item before any re-assessment. This module, through an accessible and straightforward interface, facilitates a variety of pseudonymization functions. These include number categorization, personal information masking, encryption, null value elimination, data substitution, column deletion, and data re-categorization, all aimed at streamlining and augmenting the efficiency of the data sharing review process.

Given the anticipated frequency of such review scenarios, the pseudonymization pipeline module, adhering to the "Guidelines for the Use of Medical Data" for every data item, becomes a critical asset in supporting a more streamlined and effective review process. By ensuring that each piece of data is treated with the appropriate level of privacy and security measures, the module aids in maintaining a high standard of data protection while facilitating the necessary data sharing and utilization.

D. Advancing Pseudonymization and Review Functions in the KMBIG System

The KMBIG system stands as a comprehensive Korean medicine big data platform, aggregating and administering medical data from a variety of healthcare institutions. It's engineered to promote the ethical utilization of medical data for research among other purposes, all while ensuring the confidentiality and privacy of individuals involved. Within the KMBIG framework, every data type is subjected to pseudonymization, adhering strictly to medical data usage guidelines that dictate specific pseudonymization techniques to safeguard individual privacy.

The system is equipped with a user-friendly interface allowing users to search for data sets and specify their data sharing and utilization intents. This feature is visualized in Figures 2 and 3, detailing the data set search process and providing a comprehensive view of the datasets available for application.

Once a user identifies the relevant data, as illustrated in Figure 4, they engage with the pseudonymization function, selecting options to anonymize or alter certain data aspects like decimal places, gender, age group, and other identifiable metrics. This process is designed to be dynamic, offering real-time data reflection to ensure users can verify and proceed with their data sharing requests confidently.

By incorporating these functionalities, the KMBIG system not only adheres to stringent data privacy regulations but also streamlines the data sharing process, making it more accessible and manageable for users aiming to contribute to and benefit from medical research and data analysis.

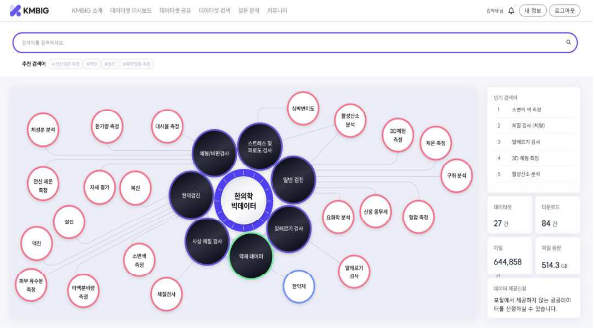
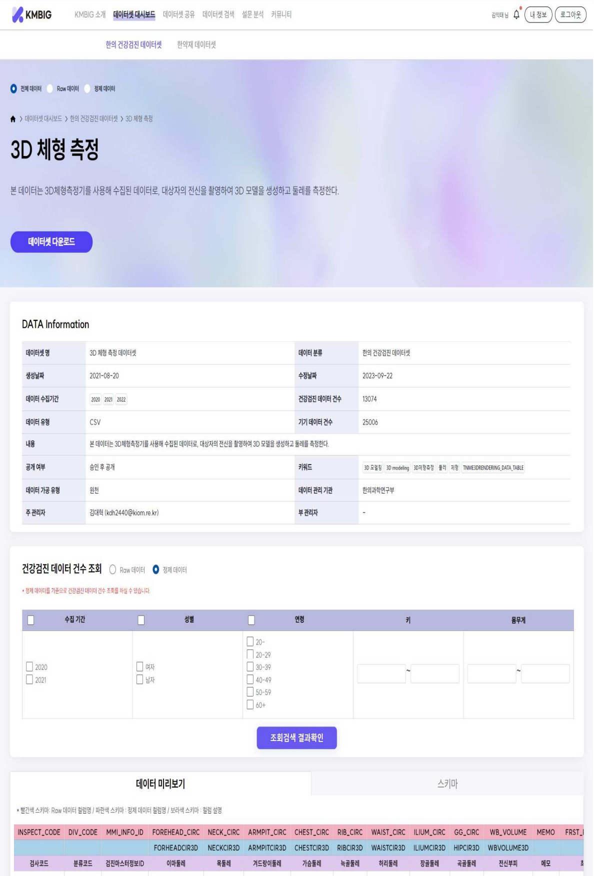


Figure 2. The Type of KMBIG Data Inforgraphic



3D 체형 측정

본 데이터는 3D체형측정기를 사용해 수집된 데이터로, 대상자의 전신을 촬영하여 3D 모델을 생성하고 둘레를 측정한다.

DATA Information

데이터셋 명	3D 체형 측정 데이터셋	데이터 분류	한의 건강검진 데이터셋
생성날짜	2021-08-20	수정날짜	2023-09-22
데이터 수집기간	2020 2021 2022	건강검진 데이터 건수	13074
데이터 유형	CSV	기기 데이터 건수	25006
내용	본 데이터는 3D체형측정기를 사용해 수집된 데이터로, 대상자의 전신을 촬영하여 3D 모델을 생성하고 둘레를 측정한다.		
공개 여부	승인 후 공개	키워드	3D 모델링 3D modeling 3D체형측정 둘레 체형 TNME3DRENDERING_DATA_TABLE
데이터 가공 유형	원천	데이터 관리 기관	한의과학연구부
주 관리자	김대혁 (kdh2440@kiom.re.kr)	부 관리자	-

건강검진 데이터 건수 조회 ○ Raw 데이터 ● 정제 데이터

* 정제 데이터를 기준으로 건강검진 데이터 건수 조회를 하실 수 있습니다.

□ 수집 기간	□ 성별	□ 연령	키	몸무게
□ 2020 □ 2021	□ 여자 □ 남자	□ 20- □ 20-29 □ 30-39 □ 40-49 □ 50-59 □ 60+	~	~

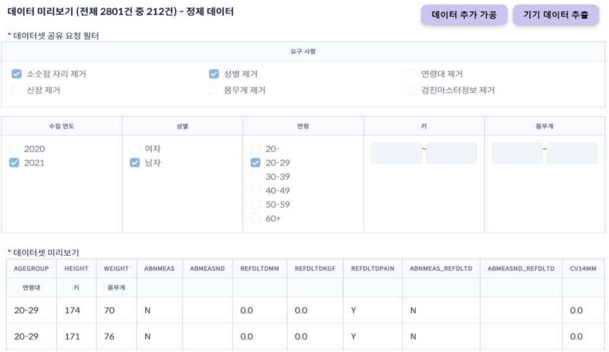
조회검색 결과확인

데이터 미리보기 | 스키마

* 빨간색 스키마: Raw 데이터 컬럼명 / 파란색 스키마 : 정제 데이터 컬럼명 / 보라색 스키마 : 컬럼 설명

INSPECT_CODE	DIV_CODE	MMI_INFO_ID	FOREHEAD_CIRC	NECK_CIRC	ARMPIT_CIRC	CHEST_CIRC	RIB_CIRC	WAIST_CIRC	ILIUM_CIRC	GG_CIRC	WB_VOLUME	MEMO	FRST_
			FORHEADCIR3D	NECKCIR3D	ARMPITCIR3D	CHESTCIR3D	RIBCIR3D	WAISTCIR3D	ILIUMCIR3D	HIPCIR3D	WBVOLUME3D		
검사코드	분류코드	검진마스터정보ID	이마둘레	목둘레	겨드랑이둘레	가슴둘레	늑골둘레	허리둘레	장골둘레	곡골둘레	전신부피	메모	초

Figure 3. Dataset Detailed View UI



데이터 미리보기 (전체 2801건 중 212건) - 정제 데이터

* 데이터셋 공유 요청 필터

요구 사항		
☑ 소숫점 자리 제거	☑ 성별 제거	□ 연령대 제거
□ 신장 제거	□ 몸무게 제거	□ 검진마스터정보 제거

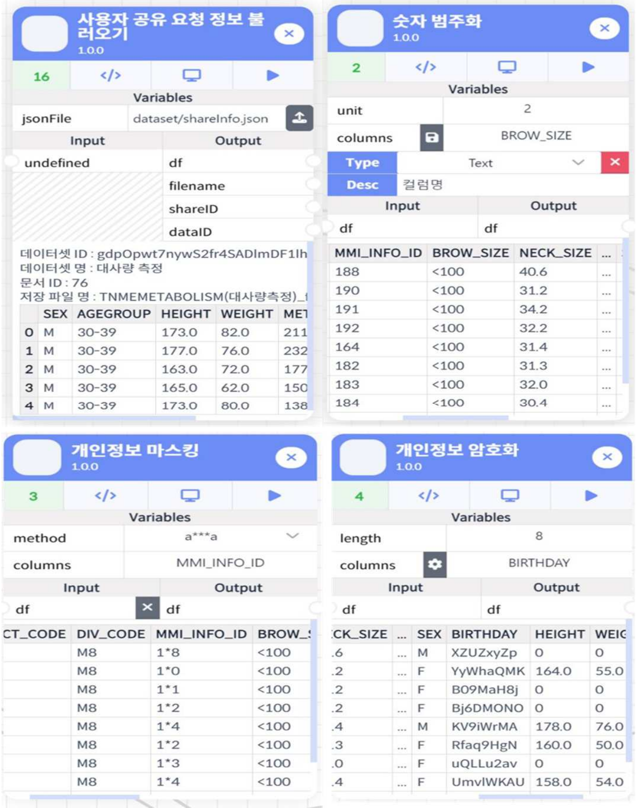
수집 연도	성별	연령	키	몸무게
□ 2020 ☑ 2021	□ 여자 ☑ 남자	□ 20- ☑ 20-29 □ 30-39 □ 40-49 □ 50-59 □ 60+	~	~

* 데이터셋 미리보기

AGEGROUP	HEIGHT	WEIGHT	ABNMEAS	ABMEASND	REFDLTDMM	REFDLTDKGF	REFDLTDPAIN	ABNMEAS_REFDLTD	ABMEASND_REFDLTD	CV14MM
연령대	키	몸무게								
20-29	174	70	N		0.0	0.0	Y	N		0.0
20-29	171	76	N		0.0	0.0	Y	N		0.0

Figure 4. De-Identification Processing UI

Figure 5 presents the digest function interface within the KMBIG system, a specialized feature that enables administrators to incorporate pseudonymization elements, encrypt data, or enhance the anonymization level. Tailored for user convenience and operational efficiency, this function streamlines the management and application of pseudonymization techniques across various data sets, furthering the system's commitment to robust data privacy.



Variables	
jsonFile	dataset/shareInfo.json

Input	Output
undefined	df
	filename
	shareID
	dataID

데이터셋 ID : gdpOpwt7nywS2fr4SADImDF1lh
데이터셋 명 : 대사량 측정
문서 ID : 76
저장 파일 명 : TNMEMETABOLISM(대사량측정)_

	SEX	AGEGROUP	HEIGHT	WEIGHT	MET
0	M	30-39	173.0	82.0	211
1	M	30-39	177.0	76.0	232
2	M	30-39	163.0	72.0	177
3	M	30-39	165.0	62.0	150
4	M	30-39	173.0	80.0	138

Variables	
unit	2
columns	BROW_SIZE
Type	Text
Desc	컬럼명

Input	Output
df	df

MMI_INFO_ID	BROW_SIZE	NECK_SIZE	...
188	<100	40.6	...
190	<100	31.2	...
191	<100	34.2	...
192	<100	32.2	...
164	<100	31.4	...
182	<100	31.3	...
183	<100	32.0	...
184	<100	30.4	...

Variables	
method	a***a
columns	MMI_INFO_ID

Input	Output
df	df

CT_CODE	DIV_CODE	MMI_INFO_ID	BROW_S
	M8	1*8	<100
	M8	1*0	<100
	M8	1*1	<100
	M8	1*2	<100
	M8	1*4	<100
	M8	1*2	<100
	M8	1*3	<100
	M8	1*4	<100

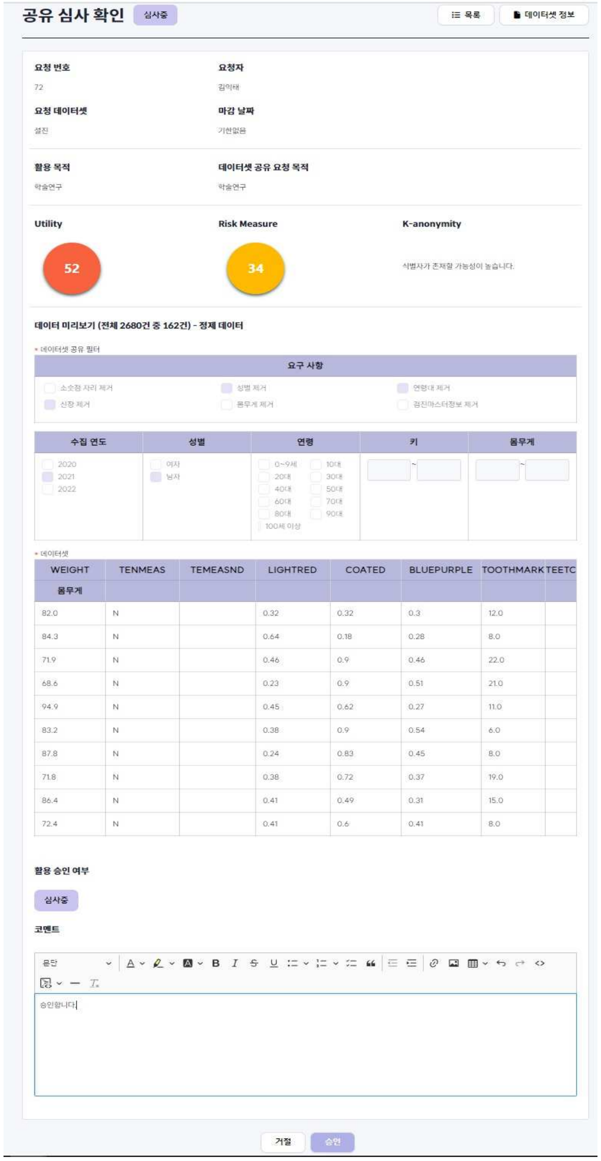
Variables	
length	8
columns	BIRTHDAY

Input	Output
df	df

CK_SIZE	...	SEX	BIRTHDAY	HEIGHT	WEIG
.6	...	M	XZUZxyZp	0	0
.2	...	F	YyWhaQMK	164.0	55.0
.2	...	F	B09MaH8j	0	0
.2	...	F	Bj6DMONO	0	0
.4	...	M	KV9iWrMA	178.0	76.0
.3	...	F	Rfaq9HgN	160.0	50.0
.0	...	F	uQLLu2av	0	0
.4	...	F	UmvlWKAU	158.0	54.0

Figure 5. Additional Pseudonymization

Figure 6 introduces the data review committee's interface, a crucial component of the system's governance. It primarily displays user data sharing applications alongside the stability report, a critical tool for evaluating re-identification risks. This report includes comprehensive details on the anonymization methods applied to each data type, the level of anonymization achieved, and the utility value of the data. By providing such granular insights, the stability report enables the committee to make well-informed decisions, assessing the balance between data utility and privacy risk, and accordingly approve, modify, or reject data sharing applications.



공유 심사 확인 심사중

요청 번호
72

요청자
김익태

요청 데이터셋
설진

마감 날짜
기한없음

활용 목적
학술연구

데이터셋 공유 요청 목적
학술연구

Utility
52

Risk Measure
34

K-anonymity
식별자가 존재할 가능성이 높습니다.

데이터 미리보기 (전체 2680건 중 162건) - 정제 데이터

* 데이터셋 공유 필터

요구 사항		
□ 소숫점 자리 제거	☑ 성별 제거	☑ 연령대 제거
☑ 신장 제거	□ 몸무게 제거	□ 검진마스터정보 제거

수집 연도	성별	연령	키	몸무게
□ 2020 ☑ 2021 □ 2022	□ 여자 ☑ 남자	□ 0~9세 □ 10대 □ 20대 □ 30대 □ 40대 □ 50대 □ 60대 □ 70대 □ 80대 □ 90대 □ 100세 이상	~	~

* 데이터셋

WEIGHT	TENMEAS	TEMEASND	LIGHTRED	COATED	BLUEPURPLE	TOOTHMARK	TEETC
몸무게							
82.0	N		0.32	0.32	0.3	12.0	
84.3	N		0.64	0.18	0.28	8.0	
71.9	N		0.46	0.9	0.46	22.0	
68.6	N		0.23	0.9	0.51	21.0	
94.9	N		0.45	0.62	0.27	11.0	
83.2	N		0.38	0.9	0.54	6.0	
87.8	N		0.24	0.83	0.45	8.0	
71.8	N		0.38	0.72	0.37	19.0	
86.4	N		0.41	0.49	0.31	15.0	
72.4	N		0.41	0.6	0.41	8.0	

활용 승인 여부
심사중

코멘트

승인합니다

Figure 6. A Safety Report Supporting the Review UI

The KMBIG system's commitment to a meticulous review process is exemplified by the generation and utilization of stability reports. These reports facilitate a quantitative assessment of re-identification probabilities, allowing for a more nuanced and effective review. This process ensures that the data sharing and utilization within the KMBIG ecosystem are conducted under the highest standards of ethical and privacy considerations.

III. Discussion and Conclusion

In this study, we explored guidelines and implementation strategies for Korean traditional medicine clinical big data, considering various aspects, in the context of the increased importance of data utilization and privacy protection technologies due to amendments in laws related to personal information protection. We proposed a systematic approach to effectively manage and securely share Korean traditional medicine clinical data by considering data types, the necessity of anonymization, risk levels, and more. Additionally, we established the review process of the Data Ethics Committee and provided post-anonymization stability reports, presenting a comprehensive system that simultaneously considers data utility and privacy protection.

For the future, we plan to conduct usability test targeting both the general public and experts to evaluate and enhance the usability of the KMBIG system. We expect that the outcomes of this research will contribute to the advancement of both data utilization and personal information protection.

Acknowledgment

This study was conducted with the support of the Korea Institute of Oriental Medicine's project titled 'Clinical Big Data Collection and Service Platform Construction for AI Traditional Medicine Doctor Development (KSN1922110)' in the year 2023. We express our gratitude to the relevant agencies for their support.

References

[1] Korea Internet & Security Agency, “2023 Privacy Report : Monthly Analysis of Personal Information Protection Trends”, Korea Internet & Security Agency, March 2023.

[2] Ministry of Public Administration, “Security Personal Information Protection Guidelines : Ministry of Public Administration and Security Ordinance No. 167”, National Law Information Center, November 2020. https://www.mois.go.kr/frt/sub/a08/personalInfo/screen.do

[3] Ministry of Culture, Sports and Tourism , “South Korea Policy Briefing : Data 3 Lows”, Ministry of Culture, Sports and Tourism, November 2021.

[4] Personal Information Protection Commission, “Information on Amendments to the Personal Information Protection Act and Enforcement Decree (Draft)”, Personal Information Protection Commission, September 2023.

[5] K-H. Lee, and K-H. Kim, “A Study on the Contents and Limitation of Guidelines for Utilization of Healthcare Data”, Institute for Law of Science & Technology, Vol. 26, No. 4, 89-118, 2020.

[6] H. W. Jung, Y. S. Cho, G. W. Ko, J-I Song, and D. H. Yu, “Comparison study of synthetic data generation methods for credit card transaction data”, Journal of the Korean Data And Information Science Society, Vol. 34, No. 1, pp. 49-72, 2023.

[7] Ministry of Health and Welfare, “Healthcare and Medical Data Utilization Guidelines”, Ministry of Health and Welfare, December 2022.

[8] J. B. Lee, "A Legal study on the scope and limitations of health · medical big data utilization," Institute for Legal Study, Vol. 45, No. 1, pp. 67-99, 2021.

[9] I. T. Kim, T. H. Kim, H. C. Jang, and Y. H. Son, “Development of a Data De-identification Pipeline and Review Support Functions for Utilizing Korean Medicine BigData”, The Society of Korean Herbal Medicine Information, Vol. 11, No. 2, pp. 111-122, 2023.

[10] Little, C., Elliot, M., Allmendinger, R. and Samani, S.S., “Generative adversarial networks for synthetic data generation: a comparative study”, arXiv preprint arXiv:2112.01925, 2021.

[11] Y. J. Choi, S. H. An, and J. Y. Bae, “Trends in Research on AI-Based Face Image De-Identification Technology”, The Journal of The Korean Institute of Communication Sciences, Vol. 39, No. 12, pp. 25-31, 2022.

[12] S. J. Yoo, and N. R. Park, “Synthetic Data Generation for Individual Credit Data Using CART”, Journal of the Korean Official Statistics, Vol. 25, No. 1, pp. 1-30, 2020.

[13] K. W. Kim, and H. K. Min. “Mobility Creation of Reproduction Data (Synthetic Data) and Statistical Verification”. Korean Society of Transportation, Vol. 20, No. 2, pp. 6-13, 2023.

[14] S. H. Ryu, Y. K. Hong, G. H. Ko, H. D. Yang, and J. W. Kim, “Privacy Model Recommendation System Based on Data Feature Analysis”. Journal of the Korea Society of Computer and Information, Vol. 28, No. 9, pp. 81-92, 2023.