Original Article
Yonsei Med J 2022 Jan;63 Suppl:S84-92
https://doi.org/10.3349/ymj.2022.63.S84
Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Concept and Proof of the Lifelog Bigdata Platform for Digital Healthcare and Precision Medicine on the Cloud

Kyu Hee Lee[1,2], Erdenebayar Urtnasan[1,2], Sangwon Hwang[1,2], Hee Young Lee[2,3], Jung Hun Lee[3], Sang Baek Koh[2,4], and Hyun Youk[2,3]

[1]Artificial Intelligence Big Data Medical Center, Yonsei University Wonju College of Medicine, Wonju;
[2]Bigdata Platform Business Group, Yonsei University Wonju College of Medicine, Wonju;
Departments of [3]Emergency Medicine and [4]Preventive Medicine, Yonsei University Wonju College of Medicine, Wonju, Korea.

Purpose: We propose the Lifelog Bigdata Platform as a sustainable digital healthcare system based on individual-centric lifelog datasets and describe the standardization of lifelog and clinical data in its full-cycle management system.
Materials and Methods: The Lifelog Bigdata Platform was developed by Yonsei Wonju Health System on the cloud to support digital healthcare and precision medicine. It consists of five core components: data acquisition system, de-identification of individual information, lifelog integration, analyzer, and service. We designed a gathering system into a dedicated virtual machine to save lifelog or clinical outcomes and established standard guidelines for maintaining the quality of gathering procedures. We used standard integration keys to integrate the lifelog and clinical data. Metadata were generated from the data warehouse after loading combined or fragmented data on it. We analyzed the de-identified lifelog and clinical data using the lifelog analyzer to prevent and manage acute and chronic diseases through providing results of statistics on analysis.
Results: The big data centers were built in four hospitals and seven companies for integrating lifelog and clinical data to develop the Lifelog Bigdata Platform. We integrated and loaded lifelog big data and clinical data for 3 years. In the first year, we uploaded 94 types of data on the platform with a total capacity of 221 GB.
Conclusion: The Lifelog Bigdata Platform is the first to combine lifelog and clinical data. The proposed standardization guidelines can be used for future platforms to achieve a virtuous cycle structure of lifelogging big data and an industrial ecosystem.

Key Words: Lifelog, big data, digital health, precision medicine

INTRODUCTION

Lifelogs are real-world data on daily life that are usually stored in personal storage or cloud storage. Lifelogging refers to the processes of data acquisition and using various sensors and smart devices.[1] The lifelog datasets for digital health consist of one's daily life data and clinical data for individuals in hospitals. A lifelog dataset is informative and powerful in developing digital healthcare services because clinical data can be included in daily life. Therefore, lifelog data are developing into novel research topics, particularly in regards to whether they can improve daily life quality and expand insights into digital health and precision medicine. However, it can be difficult to actively use lifelogged data because their fragments are stored by various entities, such as local hospitals, service providers, and individuals.

Rates of chronic disease and death have increased with increases in the aging population. The management of chronic disease is expanding from the hospital to the individual with the development of digital health and information technology.

Received: September 14, 2021 **Revised:** October 21, 2021
Accepted: November 5, 2021
Corresponding author: Hyun Youk, MD, Department of Emergency Medicine, Yonsei University Wonju College of Medicine, 20 Ilsan- ro, Wonju 26426, Korea.
Tel: 82-33-741-5401, Fax: 82-33-741-5432, E-mail: yhmentor@yonsei.ac.kr

•The authors have no potential conflicts of interest to disclose.

However, perception, treatment, and control rates are significantly low in young hypertensive patients aged 20–39 years, compared to other age groups.[2] 20.9% of diabetic patients require active treatment with glycated hemoglobin (HbA1C) above 8.0%; however, only 8.4% of them show controlled blood glucose, blood pressure, and total cholesterol.[3,4] Additionally, chronic obstructive pulmonary disease (COPD) was highly prevalent in 13.3% of adults over 40% and 28.3% of people over 65 years in 2017. The mortality rate due to chronic lower respiratory tract diseases, including COPD, ranked 8th highest in causing deaths (12.9 people per 10000) in 2018, despite showing lower mortality in well managed COPD patients.[5] Therefore, it is difficult to conclude that systems to prevent chronic diseases in Korea work well.[6] Therefore, recording clinical information and lifelog data, including lifestyle habits, and integrating them into the one big data platform could potentially help with managing chronic diseases. To do so, establishing a statistical analysis system and a full cycle management system is required.[7]

The concept of a personal health record (PHR) was initially introduced by Carl Dragstedt (1956) in the US.[8] He suggested that everyone needs a good personal health log. Since Dragstedt's electronic health record (EHR) concept, several studies or solutions have been developed for hospital-based EHR or electronic medical record (EMR) services for conveniently managing a hospital and patient care. PHRs have again gained attention with the development of information technology since 2000. Moreover, its value in digital health and precision medicine has become increasingly important with improvement of cloud-based big data integration systems. Therewith, health records in hospitals and technological advancements, such as those in wearable devices, smartphones, and the Internet of things (IoT), can be integrated with out-of-hospital health records, resulting in complete PHR and big data.[9] Complete PHRs, including lifelogs, are targets for digital health and precision medical development. However, PHR has been separated from EMR-based medical big data and lifelog big data. EMR-based medical big data platforms have been designed and implemented for healthcare services, in which the United States outperforms other countries. Europe (Austria, Denmark, England, Estonia, Finland, Norway, Portugal, and Sweden) and Australia have already applied big data platforms in clinical data from hospitals for digital healthcare.[10] The Australian digital health agency manages and operates a big data platform for healthcare services, which is operated with personally controlled EHRs.[11] In England, the department of health and social care manages and provides clinical data and digital healthcare services based on big data platforms.[12] However, none of these platforms and solutions contain lifelog data; they are based only on hospital-based data. Moreover, they focus on connections between the hospital and the individual.[13-16] Additionally, datasets stored in those platforms are fragmented into in-hospital data and out-of-hospital data; hence, they cannot be appropriate with analysis for digital healthcare services.[17] Therefore, a complete PHR dataset must be combined with the EMR dataset to complete digital health and precision medicine. Merzghani, et al.[18] proposed a semantic big data platform to integrate heterogeneous wearable data and designed an efficient architecture for heterogeneous data storage, including structured and unstructured data. Additionally, Suciu, et al.[19] proposed the architecture of a cloud computing system for big data processing and storage. They used IoT and machine-to-machine communication technologies in their platform to provide telemedicine and e-health services. Furthermore, Manogaran, et al.[20] studied the novel architecture of big data platform based on IoT for big medical data storage and processing. De-identification or anonymization should be performed according to standardized guidelines for personal and private information after performing these merging operations.[21-23]

In this study, we developed and proved the concept of the Lifelog Bigdata Platform to provide healthcare services based on the cloud. We designed and implemented a big data platform consisting of five main components: lifelog acquisition, integration, de-identification, analysis, and services. In the lifelog acquisition component, a data acquisition system (DAS) was employed to measure and upload lifelog data generated in the full cycle of an individual's life. Additionally, guidelines and standards were provided to collect and transmit lifelog data from communication systems to cloud storage. Data integration and de-identification methods were used to generate an informative and safe dataset for diversified analyses and applications. Finally, the Lifelog Bigdata Platform can provide digital healthcare services and precision medicine services for future medicine.

MATERIALS AND METHODS

Data centers acquired clinical, lifelog, and genome data and transmitted them to the proposed Lifelog Bigdata Platform. Each data center as a data provider could upload their data to a dedicated machine using a secure socket layer (SSL) virtual private network (VPN). The platform contained a web-based interface program that uploaded data using a predefined catalog of data.[24-26] The DAS on the cloud transmitted data from centers to the platform area using the developed application programming interface (API) or the agent. Furthermore, the API and agent included modules for pseudonymizing or anonymizing individual sensitive information of the DAS. Data loaded from the centers were combined with a serial dataset using JOIN keys. Moreover, combined serial data that had undergone refinement and de-identification were processed using consumer-specific significant statistics or visualized through a lifelog big data analysis system. A clinical or lifelog data model was classified based on the data type and stored in a data warehouse (DW). Furthermore, processed and analyzed data were stored in the DW based on its type (i.e., the clinical or lifelog model). Finally,

the processed data were sold for a fee or provided for free in the market based on the data pricing policy within the Lifelog Bigdata Platform.

Data centers

The Lifelog Bigdata Platform employed two types of data centers. First, medical data centers produced data from clinical processes and clinical trials in hospitals. Second, lifelog data centers generated data based on lifelog using IoT devices as hardware gateways or smartphone applications as software gateways. All data centers were physically implemented on their infrastructure, separated by the main platform. Moreover, data centers used a dedicated machine on the cloud, as indicated by the DAS (Fig. 1). DAS consisted of virtual machines (VMs) and conducted preprocessing steps, such as validation checks and de-identification, for transmitting data to the platform. We designed a DAS architecture with robust security; only officers from each data center could access it via SSL VPN.

Medical data center

Data centers producing clinical data from clinical processes and trials were implemented based on a DW. The medical data centers had a common work scope suitable for multi-institution research and data transmission and processing. They utilized cohort data from medical research to identify causal relationships between risk factors and disease occurrence by checking health conditions through long-term follow-up. In addition to previously established clinical data, lifelog data were collected using wearable devices and smartphone applications through clinical trials. This study was approved by the Institutional Review Board of Wonju Severance Christian Hospital (CR319318, CR320120, and CR320162).

Lifelog data center

Medical device companies and startups produced data, such as blood sugar and blood pressure, based on wearable devices and certified medical devices. They managed data from medical or wearable devices by transmitting them to servers designated by companies. As it could be challenging to maintain the data quality or to use standardized methods, compared to medical data centers, due to a lack of human and technical resources, the Lifelog Bigdata Platform utilizes a standardized data catalog and quality improvement policy.

Clinical and lifelog data transmission

We implemented repositories in VM to fit the standardization principle based on the data processing system as per the data management policy. Moreover, data centers produced raw data or files in dedicated VMs situated on the cloud. We considered situations where the file system, RDBMS, and NoSQL were available because data were classified into three types: structured, semi-structured, and unstructured. Therefore, we chose distributed storage and a processing method depending on how the data were loaded. The Lifelog Bigdata Platform provides step-by-step uploading processes for data loaded from physical data centers to dedicated machines on the cloud to collect data efficiently.

We developed a GUI-based user interface (called LifelogUploader) to upload data through a set of sequences. LifelogUploader validates the data based on the data catalog defined

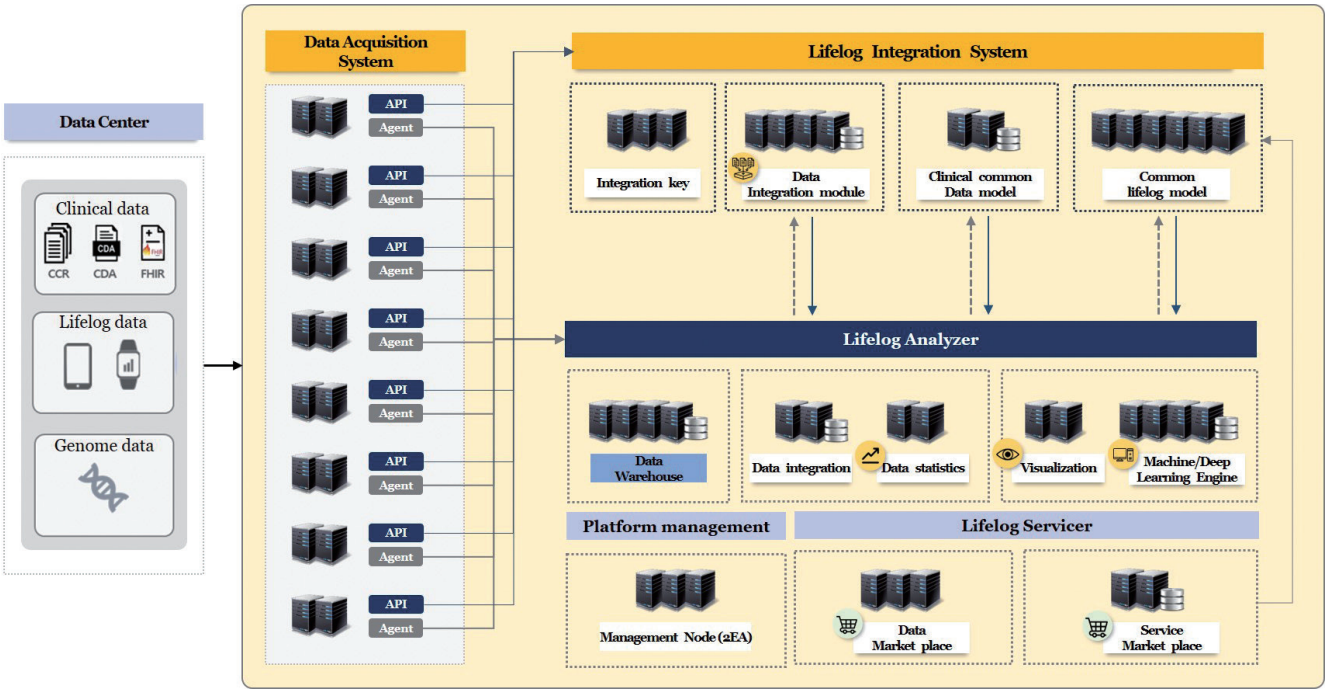

Fig. 1. The architecture of the Lifelog Bigdata Platform based on the cloud.

https://doi.org/10.3349/ymj.2022.63.S84

in each center to determine whether it is uploaded. The LifelogUploader monitors the collection status of uploaded data in real-time or periodically on the dashboard. Also, it retrieves the collection history and capacity of each object. It generates log files whenever data are not uploaded by a validation check.

Lifelog data platform

We obtained a cloud security assurance program (CSAP) to load sensitive data, including medical and personal information, on the cloud. Specifically, medical information was certified per ISO/IEC 27799, a guideline for organizational information security standards and information security management practices. The Lifelog Bigdata Platform equipped with the security authentication system consists of loading the collected data, encryption, de-identification, data processing and integration, and a data open system. Detailed information in collected data is registered and managed in a data catalog to monitor and process the acquisition and processing status. The catalog has criteria, such as registration and modification date, operator, center and data source, storage location, real-time and batch processing, and characteristic information.

Data acquisition system

Dedicated machines composed of VMs served as a guide to load the data collected and produced in the data center.[27-29] Each VM in the DAS could not be accessed by other center managers or the platform administrator; only the manager of the center could access it. The DAS defines the transmission processes and refers to the data collection history or capacity for each collection. Moreover, the DAS generates real-time notifications to warn for uncollected data and system failures; it provides de-identification of personal information. Data loading in the DAS uses a queue method; therefore, redundant processing or omission does not occur between the dedicated machines and the platform, even for server and service interruption problems. We developed a function in terms of data management to easily test the object to be deleted by setting the archive cycle and the total amount of the original data. We adopted the Integrating the Healthcare Enterprise (IHE) information technology infrastructure profile for data transmission between the DAS and lifelog integration system (LIS), the audit trail and node authentication to establish security measures, the Pull-style method for document metadata subscription, and the cross-enterprise document reliable interchange to provide document interchange. The API transmits data to the platform using a PULL method in the absence of personal information in the collected data. However, data containing sensitive information related to privacy are transmitted to the platform through pseudonymizing or anonymizing personal information.

The DAS agent is capable of controlling a workflow-based acquisition process and setting the data collection cycle. It can add and delete data collection processes in real-time or in batches. The platform has a system that manages the data lifecycle: data collection, operation, utilization, and disposal to control data quality. Furthermore, a quality management system determines the appropriateness of stored data type based on the defined table scheme by registering and managing metadata of collected data.

De-identification is performed when data with personal information are stored in the database. Otherwise, data are stored by applying encryption through hash algorithms, such as SHA-256/512, as needed. The Health Insurance Portability and Accountability Act was enacted in the United States in 1996 to standardize the electronic exchange of healthcare-related administrative and financial data. Furthermore, Korea released revised guidelines for the safe use of healthcare data in September 2020. The de-identification pseudonymizes or anonymizes personal information using the method suggested in these guidelines. We applied three techniques, the k-anonymity, l-diversity, and t-closeness, presented in the guidelines for de-identification of personal information. Before de-identified data are stored in the DW on the platform, re-identification risk is estimated through risk analysis provided by the ARX anonymization framework[30] as the evaluation tool for de-identification. The risk analysis in ARX uses three attacker models: prosecutor, journalist, and marketer. In each of the models, the re-identification risk is reported into three categories, such as records at risk, highest risk, and success rate. We focused on success rate (i.e., the proportion of records that can be re-identified on average) and transmitted data to the DW if the success rate was lower than 0.01%.

Lifelog integration system

An LIS created a standardized model to manage and process common data after verifying the re-identification algorithm of de-identified data. Statistical data or analysis data generated by data processing are stored in the DW within the Lifelog Bigdata Platform. Pseudonymized data can be combined with additional information for data integration. The platform's authentication server generates a JOIN key for data convergence and then combines its information from the same person through the combination key management agency and the data combination agency. Additional information on individuals used in data combination is stored and managed in a physically separate place from de-identified data at the time of pseudonymization.

The functions of the LIS consist of data processing and integration, transmission within the platform, data management using metadata and schema, and quality management. The LIS can quickly store and search structured data, including time-series data in the form of continuous numerical data, and uses an in-memory search engine to quickly identify and assemble necessary contents in a large amount of unstructured textual data. Specifically, we applied file system-based technologies, such as Hadoop, to mass process large unstructured data, such as images and videos. The data processing step supports extract, transform, and load (ETL) functions. Additionally, it sup-

ports data presets to manage and change uploaded data into the desired set. Combinations and links to various datasets include outlier detection, replacement, deletion, and row filtering by data cleansing tasks and conditions and substitution of regular expressions. The ETL engine performs batch operations and scheduling for large-scale data processing after completing the data processing. Additionally, the data are stored in a specified file system or Hadoop distributed file system.

Data management utilizes the schema and version checking technique of the original data. The data distribution is verified through a preview screen for the connected dataset. Metadata are managed according to data name, description, keyword, download path, URL, and registration date. The schema is managed according to its name, information, and description. Metadata allows users to automatically extract their data using change management, patches, and data collection items. Therefore, it is possible to automatically extract metadata from the owned DB and store it in another DB or download it in various formats, such as CSV, XLSX, and TXT. Additionally, standard metadata are selected through an automatic recommendation function of the system; however, it registers metadata in advance in the absence of an appropriate standardization scheme.

Data quality management and verification are handled by deriving and managing elements, such as the data quality index, critical to quality, and business rules. Each center performs the necessary analysis and inspection by utilizing the structure and collection status to store the generated data. Therefore, the Lifelog Bigdata Platform establishes its verification system based on the public information quality management manual for data standardization and quality improvement. We adopted a standard word dictionary, standard terminology, column definition, domain definition, and table definition to increase quality management efficiency. For managing data quality, we use a relational database checker (RDBChecker) program, which was developed by the regulatory agency. The program verifies data consistency, referential integrity, and entity integrity from the database on the platform and estimates the defect ratio of data and Six-sigma, which is a quality management methodology developed by Motorola, Inc. in 1986. This approach uses data-driven reviews to limit mistakes or defects in enterprises or business processes. Moreover, Six-sigma, a six-standard deviation event from the mean, is required for a mathematical error. We defined a standard data dictionary that complied with the standard terms of the government data platform. Furthermore, the data platform-based linkage through the standardization system was designed by introducing an open source-based data platform, such as CKAN, to link data easily.

Lifelog analyzer

The lifelog analyzer was built using representative open-source tools, such as R-Studio, Zeppelin, and Jupyter. This system analyzes the loaded lifelog and clinical data using the constructed statistical package and analyzes clinically significant data related to chronic diseases. The results of data analysis derived using the analysis system are fetched from the DW. After that, significant data resulting from deep or machine learning are again stored in the DW for data distribution. Moreover, we also used a machine learning and deep learning engine to provide a full learning pipeline for preprocessing, model design, result management, distribution, and inference to derive robust learning processes and visualize results. The system generates new data by analyzing datasets tailored to companies or researchers and significant data to be used to prevent and analyze each disease. The raw data repository of the platform can be accessed and analyzed through a given interface only. We implemented a workflow design function redefined by a system administrator to upgrade the analyzed or derived public datasets continuously. This allows the workflow to manage repetitive data processing and monitor the processing progress and results.

Lifelog services

Lifelog services consist of data and service markets to provide clinical and lifelog data. A data market is a virtual place where the lifelog datasets are presented as products from the Lifelog Bigdata Platform to platform users. A service market is a virtual place where free and paid services are released, including data analysis services, data integration services, and innovation services supported by the Lifelog Bigdata Platform. In addition, the service system contains a data distribution process (i.e., preprocessing, integration, and analysis) in the DW; the data product is prepared through this process. Hence, data integration and data visualization are important for improving the value and use of datasets. These processes and functions on the lifelog services were implemented using the CKAN framework. The backend and frontend services that support the data service system are summarized in Table 1.

Policies

The proposed Lifelog Bigdata Platform promotes several lifelog data standardization policies, data distribution, and data privacy/security. First, we standardized the lifelog big data for data acquisition and storage on the platform or cloud system. Lifelog data standardization was performed using the standard glossary, specification, or registry for the table, column, and standard

Table 1. Functions of the Backend and Frontend of the Data Service System

Lifelog provider	Lifelog consumer
Product upload management	Login (authorization)
Product definition	Recent data overview
Preview, detailed view, access control	Data catalog
Data lifecycle management	Product manage, list, category,
Log and statistics	Visualize, search, external search
User management	User manual
Purchase request and approve	Purchase and payment
Data download link generation	Data download and log view

https://doi.org/10.3349/ymj.2022.63.S84

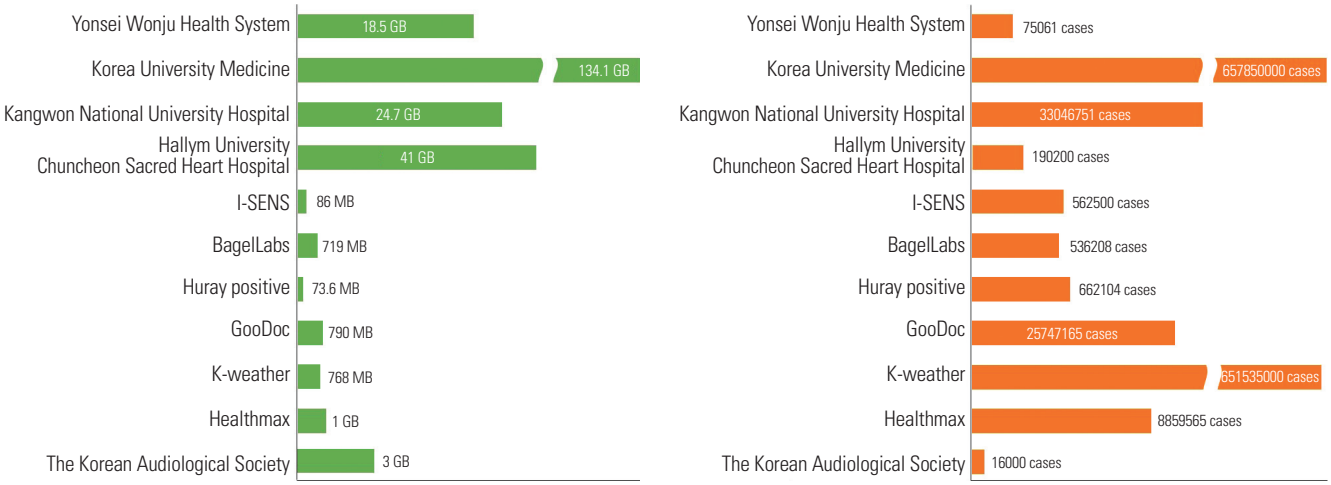

Fig. 2. Data representing capacity and the number of cases in the repository on the Lifelog Bigdata Platform.

operating procedure. Next, policies were made for the data distribution based on the healthcare big data application guidelines and the selling of big data. The policies cover the entire data distribution process from data acquisition to data selling throughout the lifelog big data platform. Finally, security policies were devised to cover general safety and personal privacy issues. The Lifelog Bigdata Platform was implemented on the NHN Toast cloud system certified by ISO/IEC 27799 for medical data storage and CSAP for cloud security certification. A personal privacy policy prevents or protects leakage of personal information based on personal information privacy and relevant health laws.

RESULTS

Lifelog datasets

The proposed Lifelog Bigdata Platform was opened. Additionally, some lifelog data were acquired from the data centers and stored on the platform. Fig. 2 shows that we currently have a dataset with 94 types of data, 718669989 pieces, and approximately 220 GB capacity. These lifelog data were collected from the data centers, including four medical data centers and seven lifelog data centers; their details are provided in Table 2. We evaluated Six-sigma and defect ratio of data using the RDB-Checker program for data quality assessment. In result, we obtained defect rates from 0.000000514% to 7.09% and Six-sigma values from 2.97 to 6.83 for the dataset stored in the platform, as shown in Table 3.

Lifelog innovation services

We have provided two healthcare services based on lifelog data to promote the platform and an example of the lifelogging data application. The first service was blood glucose monitoring for diabetes. We developed a prediction algorithm for diabetes monitoring using the open dataset of the Lifelog Bigdata Plat-

Table 2. Data Sources of the Lifelog Bigdata Platform

Data centers	Data sources
Yonsei Wonju Health System	Metabolic syndrome's lifelog data
	12-lead ECG data
	Cohort study data
Korea University Medicine	CDM data
	inPHR data
Kangwon National University Hospital	Lifelog data
	Clinical information data
	Clinical support data
	Health insurance and other data
	Clinic and lifelog data of newcomers
	Meal image data
Hallym University Chuncheon Sacred Heart Hospital	Smart health data in Kangwon (2014–present)
	Healthy life data in Inje-Yangu (2015–present)
	Healthy life data in Seoul (2018–present)
	Chatbot data for dementia (2018–present)
The Korean Audiological Society	Auditory test data
BagelLabs	Morphotype data
	Morphotype analysis data
Huray positive	Self-recorded data
	Intervention data
GoodDoc	Medical service data
	Registry service data
	Medical consulting data
	Insurance service data
K-weather	Life-air data for houses
	Life-air data for schools
	Life-air data for crowd facilities
	Health environment index
	Clinical trials in Wonju
	Lifelog data of vulnerable
I-SENS	Chronic disease analysis data
Healthmax	Metabolic syndrome data

form. The prediction algorithm used simple vital signs, such as age, body mass index, and glucose level; it then predicted glycated hemoglobin and total cholesterol levels using machine learning methods. The second was a blood component prediction service using an electrocardiogram (ECG) signal; it could easily and quickly predict hyperkalemia based on artificial intelligence without a blood test. A single-lead ECG signal from the wearable device was input and preprocessed. The pretrained deep learning algorithm was applied on it to predict the serum potassium level, as shown in Fig. 3.

DISCUSSION

In this study, we introduced and described the architecture and functions of the Lifelog Bigdata Platform. The platform is composed of subsystems for data acquisition, data uploading, and marketplace with in-hospital and out-of-hospital data. Using the platform, we collected 94 types of datasets loaded from four medical data centers and seven lifelog data centers with a total size of 220 GB as a part of the first-year goal. According to set guidelines, all datasets without any specific problems were finally released to the data market of the Lifelog Bigdata Platform. Some datasets were provided for free, whereas some were sold through the platform according to pricing policies.

We considered four characteristics when designing and implementing the Lifelog Bigdata Platform: cooperative, circulatory, connectable, and convergable. A cooperative platform means that the open platform can combine various datasets from various institutions, such as medical institutions, healthcare industries, and startups. A circulatory platform can provide feedback from lifelogging users to data centers through the Lifelog Bigdata Platform, and this feedback can help the officers of individual data centers generate specialized datasets to meet the requirements of users. A connectable platform can be easily extended to other big data platforms, data centers, and institutions that need to transmit bi-directionally and store metadata or analytics results. Finally, it is important that the Lifelog Bigdata Platform be convergent. A convergent platform can incorporate advanced technologies, such as wearable and smart devices, edge computing, IoT, and cloud computing; moreover, it can be integrated with clinical data and lifelog data using outstanding technologies.

Table 3. Class of Data Quality by Six-Sigma

Data centers	Process sigma	Defects (%)
Wonju Yonsei Medical Center	5.73	0.001177
Korea University Medical Center	6.83	0.0000000483
Kangwon National University Hospital	5.50	0.001771
Hallym University Chuncheon Sacred Heart Hospital	4.06	0.0512666
BagelLabs	3.77	1.15
Huray positive	2.97	7.09
GoodDoc	6.39	0.0000000514
I-SENS	5.85	0.00000687

In previous studies, healthcare platforms that could collect big data from wearable devices were based on API.[18] Furthermore, a semantic big data platform was implemented to process heterogeneous wearable data and visualize analyzed data.[19] Unfortunately, there are no standardized guidelines to collect data from multiple institutions. A secured smart healthcare monitoring and alerting system was proposed to process and analyze big data for finding valuable information.[20] The grouping and choosing system was used for securing integration on cloud computing. However, the architecture does not support any methods for privacy preservation or de-identification. To address these issues, our study designed and implemented the Lifelog Bigdata Platform to encompass all data industry stages: DAS, LIS, and data marketplace. The DAS uses agents based on VMs to acquire lifelogged datasets from medical and lifelog data centers. Data acquisition agents perform multiple tasks, including data preprocessing, validation, de-identification, and transmission to the platform. The LIS functions in managing and operating the metadata, data schema, and mapping table to generate a novel and informative lifelog dataset. Finally, all datasets throughout the acquisition and integration systems are displayed in the data marketplace as a final product. The data marketplace can perform essential tasks in the shopping mall system, such as dataset searching, overview, and purchase.

The Lifelog Bigdata Platform can acquire and establish volumes of big data using the lifelog data generated from the full cycle. Established lifelog big data can be analyzed using artificial

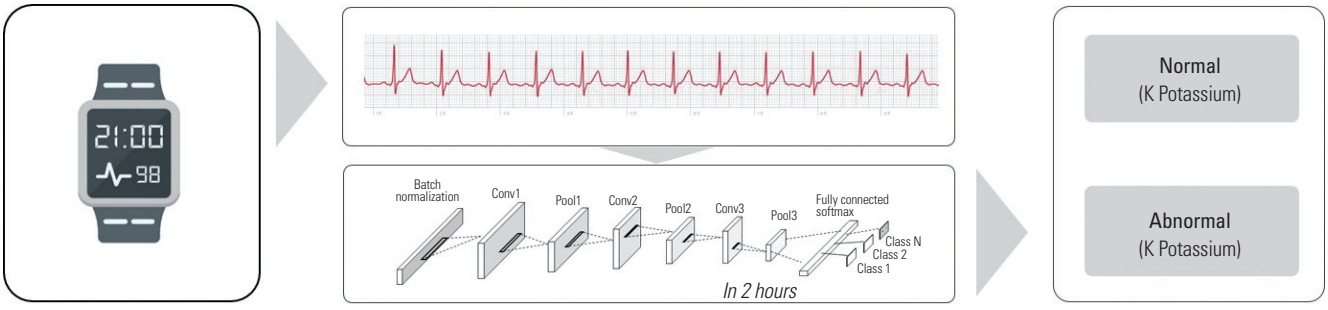

Fig. 3. Electrocardiogram-based prediction algorithm for blood components.

https://doi.org/10.3349/ymj.2022.63.S84

intelligence and deep learning for knowledge discovery, value creation, innovation services, and new business discovery. For instance, it can provide the infrastructure and environment for novel digital healthcare services, including health monitoring, diagnostic and prediction services, and comorbidity prediction. However, there are some legal and systematic issues regarding the lifelog big data platform in terms of privacy of data providers or data integration. Systemic reform can be started by promoting a new act called "three data-related law." This would help solve other issues soon. Lifelog Bigdata Platform can acquire and store more valuable and applicable data through this reform. It would provide and promote the new value and new model of the data industry.

In this study, we summarize the Lifelog Bigdata Platform based on cloud computing. In the platform, five main subsystems are employed: lifelog data acquisition, data integration, de-identification, data analysis, and data services. In summary, the DAS measures and uploads lifelog data following guidelines and standards for lifelogging data collection and transmission from communication systems to cloud storage. The data integration and de-identification methods provide informative and safe dataset generation for diversified analyses and applications. Finally, the Lifelog Bigdata Platform provides healthcare services and precision medicine services for acute and chronic diseases. This study has a limitation in that a standardized method such as HL7 has only been partially applied to produce, refine, and analyze real world data and to increase interoperability. Therefore, it is necessary to study a method for applying the messaging standard presented by HL7 or IHE to real world data in the future.

ACKNOWLEDGEMENTS

This research was supported by the National Information Society Agency (NIA) funded by the Ministry of Science, ICT through the Big Data Platform and Center Construction Project (No. 2020-Data-W123).

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I-1A1A01066463).

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C1035).

AUTHOR CONTRIBUTIONS

Conceptualization: Kyu Hee Lee, Erdenebayar Urtnasan, Sangwon Hwang, and Hyun Youk. Data curation: Kyu Hee Lee, Erdenebayar Urtnasan, and Jung Hun Lee. Formal analysis: Kyu Hee Lee, Erdenebayar Urtnasan, and Jung Hun Lee. Funding acquisition: Kyu Hee Lee, Sangwon Hwang, and Sang Baek Koh. Investigation: Kyu Hee Lee, Sangwon Hwang, Sang Baek Koh, and Hyun Youk. Methodology: Kyu Hee Lee, Erdenebayar Urtnasan, Sangwon Hwang, and Hyun Youk. Project administration: Kyu Hee Lee, Hee Young Lee, and Hyun Youk. Resources: Kyu Hee Lee and Hyun Youk. Software: Kyu Hee Lee, Erdenebayar Urtnasan, and Sangwon Hwang. Supervision: Sang Baek Koh and Hyun Youk. Validation: Sang Baek Koh and Hyun Youk. Visualization: Sang Baek Koh and Hyun Youk. Writing—original draft: Kyu Hee Lee and Erdenebayar Urtnasan. Writing—review & editing: Kyu Hee Lee, Sang Baek Koh, and Hyun Youk. Approval of final manuscript: all authors.

ORCID iDs

Kyu Hee Lee https://orcid.org/0000-0002-7378-3697
Erdenebayar Urtnasan https://orcid.org/0000-0002-3493-9724
Sangwon Hwang https://orcid.org/0000-0001-8666-7479
Hee Young Lee https://orcid.org/0000-0003-3254-2261
Jung Hun Lee https://orcid.org/0000-0002-4799-6277
Sang Baek Koh https://orcid.org/0000-0001-5609-6521
Hyun Youk https://orcid.org/0000-0002-4631-1504

REFERENCES

1. Le NK, Nguyen DH, Hoang TH, Nguyen TA, Truong TD, Dinh DT, et al. Smart lifelog retrieval system with habit-based concepts and moment visualization. Proceedings of the ACM Workshop on Lifelog Search Challenge; 2019 Jun 10-13; Ottawa: Canada; 2019. p.1-6.
2. Won JC, Lee JH, Kim JH, Kang ES, Won KC, Kim DJ, et al. Diabetes fact sheet in Korea, 2016: an appraisal of current status. Diabetes Metab J 2018;42:415-24.
3. Fisher NDL, Curfman G. Hypertension—A public health challenge of global proportions. JAMA 2018;320:1757-9.
4. Kim HC, Cho SMJ, Lee H, Lee HH, Baek J, Heo JE. Korea hypertension fact sheet 2020: analysis of nationwide population-based data. Clin Hypertens 2021;27:8.
5. An TJ, Yoon HK. Prevalence and socioeconomic burden of chronic obstructive pulmonary disease. J Korean Med Assoc 2018;61:533-8.
6. Halpin HA, Morales-Suárez-Varela MM, Martin-Moreno JM. Chronic disease prevention and the new public health. Public Health Rev 2010;32:120-54.
7. Dang LM, Piran M, Han D, Min K, Moon H. A survey on internet of things and cloud computing for healthcare. Electronics 2019;8:768.
8. Dragstedt CA. Personal health log: guest editorial. JAMA 1956;160:1320.
9. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of big data in healthcare: an overview of the European initiatives. Eur J Public Health 2019;29:23-7.
10. Salas-Vega S, Haimann A, Mossialos E. Big data and health care: challenges and opportunities for coordinated policy development in the EU. Health Syst Reform 2015;1:285-300.
11. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019;6:54.
12. Best J. The NHS App: opening the NHS's new digital "front door" to the private sector. BMJ 2019;367:l6210.
13. Park SH, Do KH, Choi JI, Sim JS, Eo H, Woo H, et al. Principles for evaluating the clinical implementation of novel digital healthcare devices. J Korean Med Assoc 2018;61:765-75.
14. Roth JA, Battegay M, Juchler F, Vogt JE, Widmer AF. Introduction to machine learning in digital healthcare epidemiology. Infect Control Hosp Epidemiol 2018;39:1457-62.
15. Jung EY, Kim J, Chung KY, Park DK. Mobile healthcare application

with EMR interoperability for diabetes patients. Cluster Comput 2014;17:871-80.
16. Ruotsalainen P. A cross-platform model for secure electronic health record communication. Int J Med Inform 2004;73:291-5.
17. Chen TS, Liu CH, Chen TL, Chen CS, Bau JG, Lin TC. Secure dynamic access control scheme of PHR in cloud computing. J Med Syst 2012;36:4005-20.
18. Mezghani E, Exposito E, Drira K, Da Silveira M, Pruski C. A semantic big data platform for integrating heterogeneous wearable data in healthcare. J Med Syst 2015;39:185.
19. Suciu G, Suciu V, Martian A, Craciunescu R, Vulpe A, Marcu I, et al. Big data, internet of things and cloud convergence–an architecture for secure e-health applications. J Med Syst 2015;39:141.
20. Manogaran G, Varatharajan R, Lopez D, Kumar PM, Sundarasekar R, Thota C. A new architecture of internet of things and big data ecosystem for secured smart healthcare monitoring and alerting system. Future Gener Comput Syst 2018;82:375-87.
21. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015; 58:S47-52.
22. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform 2017;75S:S34-42.
23. Catelli R, Casola V, De Pietro G, Fujita H, Esposito M. Combining contextualized word representation and sub-document level analysis through Bi-LSTM+ CRF architecture for clinical de-identification. Knowl Based Syst 2021;213:106649.
24. Hendriks MR, Al MJ, Bleijlevens MH, van Haastregt JC, Crebolder HF, van Eijk JT, et al. Continuous versus intermittent data collection of health care utilization. Med Decis Making 2013;33:998-1008.
25. Kumari A, Kumar V, Abbasi MY, Kumari S, Chaudhary P, Chen CM. CSEF: cloud-based secure and efficient framework for smart medical system using ecc. IEEE Access 2020;8:107838-52.
26. Shini SG, Thomas T, Chithraranjan K. Cloud based medical image exchange-security challenges. Procedia Eng 2012;38:3454-61.
27. Kim JW, Lim JH, Moon SM, Jang B. Collecting health lifelog data from smartwatch users in a privacy-preserving manner. IEEE Trans Consum Electron 2019;65:369-78.
28. Dobbins C, Fairclough S, Lisboa P, Navarro FFG. A lifelogging platform towards detecting negative emotions in everyday life using wearable devices. Proceedings of the 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); 2018 Mar 19-23; Athens, Greece: IEEE; 2018. p.206-11.
29. Kim S, Yeom S, Kwon OJ, Shin D, Shin D. Ubiquitous healthcare system for analysis of chronic patients' biological and lifelog data. IEEE Access 2018;6:8909-15.
30. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: the ARX data anonymization tool. In: Gkoulalas-Divanis A, Loukides G, editors. Medical data privacy handbook. Cham: Springer; 2015. p.111-48.

https://doi.org/10.3349/ymj.2022.63.S84