Special Article

The Cancer Clinical Library Database (CCLD) from the Korea-Clinical Data Utilization Network for Research Excellence (K-CURE) Project

Sangwon Lee¹, Yeon Ho Choi¹, Hak Min Kim¹, Min Ah Hong¹, Phillip Park¹, In Hae Kwak¹, Ye Ji Kang¹, Kui Son Choi¹,², Hyun-Joo Kong¹, Hyosung Cha¹,², Hyun-Jin Kim¹, Kwang Sun Ryu¹, Young Sang Jeon³, Hwanhee Kim³, Jip Min Jung³, Jeong-Soo Im¹,⁴, Heejung Chae¹,⁵

¹National Cancer Control Institute, National Cancer Center, Goyang, ²Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, ³Division of Data Promotion, Korea Health Information Service, Seoul, ⁴Department of Preventive Medicine, Gachon University College of Medicine, Incheon, ⁵Center for Breast Cancer, Research Institute and Hospital, National Cancer Center, Goyang, Korea

Corresponding author Heejung Chae

National Cancer Control Institute, National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang 10408, Korea
Tel: 82-31-920-1716 Fax: 82-31-920-2799 E-mail: hchae21@ncc.re.kr

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi:10.4143/crt.2024.218
Abstract

The common data model (CDM) has found widespread application in healthcare studies, but its utilization in cancer research has been limited. This article describes the development and implementation strategy for Cancer Clinical Library Databases (CCLDs), which are standardized cancer-specific databases established under the Korea-Clinical Data Utilization Network for Research Excellence (K-CURE) project by the Korean Ministry of Health and Welfare. Fifteen leading hospitals and fourteen academic associations in Korea are engaged in constructing CCLDs for 10 primary cancer types. For each cancer type-specific CCLD, cancer data experts determine key clinical data items essential for cancer research, standardize these items across cancer types, and create a standardized schema. Comprehensive clinical records covering diagnosis, treatment, and outcomes, with annual updates, are collected for each cancer patient in the target population, and quality control is based on six-sigma standards. To protect patient privacy, CCLDs follow stringent data security guidelines by pseudonymizing personal identification information and operating within a closed analysis environment. Researchers can apply for access to CCLD data through the K-CURE portal, which is subject to Institutional Review Board and Data Review Board approval. The CCLD is considered a pioneering standardized cancer-specific database, significantly representing Korea’s cancer data. It is expected to overcome limitations of previous CDMs and provide a valuable resource for multicenter cancer research in Korea.

Keywords
Common data model, Clinical data warehouse, Multicenter study, Healthcare data, Neoplasms, Database
Introduction

In recent years, the number of hospitals implementing clinical data warehouses (CDWs) has been steadily increasing. A CDW is a repository that stores and manages vast amounts of clinical data collected from various sources within a healthcare organization [1]. CDWs play a crucial role in healthcare research by enabling comprehensive data integration, standardization, and analysis. Moreover, several common data models (CDMs) have been developed to facilitate the effective utilization of data on a multicenter basis. These CDMs provide standardized database schemas that can be applied to healthcare data from diverse organizations with varying structures. For example, the Observational Medical Outcomes Partnership (OMOP) CDM is widely adopted to standardize and facilitate efficient large-scale analysis of medical data [2].

Despite the recent advances in the OMOP CDM and its extensive coverage of healthcare domains, it has limitations in the field of cancer research [3]. It still lacks a comprehensive data model specifically designed for cancer that can effectively capture cancer stage at diagnosis, treatment, and outcomes [4]. In addition, the processing of CDW data poses a major challenge for researchers, as CDWs are typically designed to minimize information loss from source data rather than to provide data tailored to the needs of back-end users. The significant time and effort required for data preprocessing are increasingly problematic in the era of machine learning, which demands access to a substantial volume of diverse clinical data.

In this context, the Korean Ministry of Health and Welfare launched the Korea-Clinical Data Utilization Network for Research Excellence (K-CURE) project in 2022, with the primary objectives of establishing cancer-specific common databases called Cancer Clinical Library Databases (CCLDs) and promoting collaborative multicenter studies. Fifteen leading hospitals in Korea are participating in the K-CURE project to establish standardized CCLDs (Fig. 1).
These hospitals are in the process of establishing in-hospital CCLDs for the following 10 primary cancer types: breast, gastric, colorectal, liver, lung, prostate, pancreatic, kidney, cervical, and hematological malignancies.

The K-CURE CCLD steering committee, comprising representatives from the 15 hospitals, is responsible for project governance, including determining essential project policies and encouraging collaborative efforts. It has established two committees to handle practical project operations—the Cancer Data Standards and Quality Committee is tasked with defining clinical data items and maintaining quality control, while the Cancer Data Construction and Utilization Committee is entrusted with enhancing clinical data accessibility and utilization.

CCLD Establishment

1. Plan for CCLD construction

The K-CURE project plans to establish CCLDs for 10 primary cancer types from 2022 to 2025, with breast and gastric cancer in 2022, colorectal and liver cancer in 2023, lung, pancreatic, and prostate cancer in 2024, and kidney, cervical, and hematological malignancies in 2025. Each year, participating hospitals choose at least one cancer-specific CCLD to establish. As of 2023, the CCLDs for breast, gastric, colorectal, and liver cancer have been completed. The CCLDs for the remaining cancer types will be completed by 2025. Updated information on the CCLD establishment status by hospital is available at the following link: https://k-cure.mohw.go.kr/.

2. Population

Each CCLD for a specific cancer type comprises patients diagnosed with that cancer
after 2010, as per the National Cancer Registration Program, with the primary tumor site classified according to the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) [5]. For example, the CCLDs for gastric cancer include patients who were initially registered in a hospital’s cancer registry under code C16.x after 2010.

3. Selection and standardization of data items

Selection and standardization of data items are performed in two steps (Fig. 2). The first step, which was completed before the initiation of the K-CURE project, aimed to determine the data necessary for cancer research. Cancer experts from diverse academic societies/groups participated in this process, determining the key data items for the 10 cancer types and assessing their clinical importance and collectability. The National Cancer Data Center (NCDC) standardized the data items and profiles across the cancer types and published “Standard Item Definition Guidelines for Cancer” for each cancer type.

The second step involves determining which data items selected in step 1 can be extracted from real-world hospital records and designing a standard schema for each cancer type-specific CCLD. A Cancer Data Standards and Quality Committee is established for each cancer type containing at least one cancer doctor, one data scientist, and one health information manager from each participating hospital. The committee identifies extractable data from each hospital and discusses the standard schema for the CCLD from both a medical and data science perspective. Subsequently, the NCDC finalizes the standard schema for the CCLD. It then distributes the schema to the hospitals, who collect the clinical data on the patients. Quality control measures for CCLDs are in accordance with six-sigma principles [6].

Korean Cancer Association
This article is protected by copyright. All rights reserved.
4. Schema design and data collection

The CCLD data frame comprises an extensive set of tables and columns in which each patient’s comprehensive cancer-related clinical data are collected throughout their entire diagnosis and treatment process (Fig. 3). The tables and columns comprise five types of clinical information: patient and tumor information, clinical assessment, test result, treatment, and follow-up (Fig. 4). For example, the CCLDs for breast and gastric cancers contain 23 tables with 513 columns and 22 tables with 498 columns, respectively. Most tables and columns are identical across all cancer types, but some are specific to a particular primary cancer type to reflect its unique features (Table 1). For instance, breast and gastric cancer CCLDs share tables for cancer characteristics and patient demographics, medical and family history, physical measurements, diagnostic information, laboratory and imaging tests, pathology results, treatment specifics, and follow-up data. Meanwhile, details on Helicobacter pylori infection and endoscopic submucosal dissection are exclusively collected in gastric cancer CCLDs, and tables for obstetric history, germline genetic tests, and gene expression assays are exclusively collected in breast cancer CCLDs. Each hospital establishes cancer-specific CCLDs with its own clinical data.

Subsequent to the establishment of a CCLD, hospitals not only update clinical data for the initial target population but also register newly diagnosed cancer patients on an annual basis. (Fig. 5). For example, the breast and gastric CCLDs were established in 2022 and initially contained data from patients diagnosed with cancer from January 1, 2010, to December 31, 2021. Then, in 2023, the CCLDs were updated to include patients newly diagnosed with cancer in 2022. Additionally, clinical data from 2022 for patients already registered in the CCLDs were added.
Data Security

CCLD data are pseudonymized to protect the privacy of patients and prevent potential risk of re-identification of individuals. In compliance with Korea’s Privacy Act, sharing personal information with a third party that could re-identify individuals from pseudonymized data is strictly prohibited. Consequently, the personal identifiable information of all patients included in a CCLD is de-identified by replacing names and social IDs with random characters. In addition, details regarding residential locations are excluded, and a patient’s date of birth is masked to contain only the year and month. Furthermore, the Ministry of Health and Welfare has published guidelines for the additional pseudonymization of medical data that can possibly re-identify individuals in the absence of personal identifiable information [7,8]. The K-CURE project also complies with these guidelines when providing CCLD data.

In addition, researchers can only access a hospital’s CCLD after securing approval from both the Institutional Review Board (IRB) and DRB of that hospital, ensuring adherence to ethical and privacy standards. Once a researcher has secured approval, a data engineer curates the database to the study’s specific population as defined in the research protocol. This data provisioning, which involves data filtering, customization, and combination, is conducted in a secure, restricted area to minimize any risk of patient privacy breaches. Analysis of CCLD data is permitted exclusively in designated secure locations reinforcing the commitment to data integrity and confidentiality. Additionally, once all analyses are complete, researchers can only retrieve the analysis results after undergoing another review to ensure the results do not contain any information that could potentially re-identify individuals in the study population.

Data Resource Access

Through the K-Cure project, researchers are able to access CCLDs from multiple
hospitals for their studies (Fig. 6). (1) Researchers visit the K-CURE portal and explore summary statistics of and metadata about the CCLDs established by each hospital and decide which CCLDs to use. The CCLD summary statistics include details about the target population size and crucial clinical data, such as diagnosis, treatment, and pathologic information. Information is available at different levels: the general public, defined as anyone who visits the portal, can assess the basic demographic information such as age and sex; researchers who wish to use a CCLD for a study can access more detailed statistical summaries relevant to cancer research. (2) Researchers recruit collaborators from the candidate institutions. (3) Subsequently, researchers must obtain approval by the IRB and the DRB of the respective hospitals to ensure the planned study is unlikely to raise ethics or privacy issues. (4) These approvals are submitted to the K-CURE portal. (5) Data engineers from each hospital then extract customized data from their CCLD and transfer them to a Central Data Center which is a physically secure environment with a protected network. (6) Finally, researchers can access and analyze the data at the Central Data Center. Once researchers complete their analyses, they are only allowed to retrieve their results, preventing any risk of re-identification.

**Strengths and Limitations**

The K-CURE CCLDs are a significant representation of South Korea’s cancer data. The fifteen major hospitals participating in the K-CURE project are pioneers in cancer care in the country. The Health Insurance Review and Assessment Service provides well-established guidelines for cancer diagnosis and treatment, consistently monitoring healthcare services to ensure they follow uniform standards, as the nation’s obligatory National Health Insurance covers the entire population [9-12]. Consequently, all hospitals consistently adhere to these guidelines. Collaborations among prominent, high-volume hospitals and the uniform nature of the Korean
healthcare system ensure the representativeness of K-CURE CCLDs, which have great potential to continue building Korea’s cancer care research capabilities and improving outcomes. The CCLD summary statistics available at the K-CURE portal aid researchers in assessing and identifying the datasets most suitable for their study. Additionally, because CCLDs are standardized and designed for a specific cancer type, researchers can minimize the amount of time they spend preprocessing and consolidating data. Finally, allowing access to CCLD data exclusively in a centralized secure environment reduces the need for extensive pseudonymization. When researchers consolidate multi-hospital data outside the K-CURE system, they receive heavily pseudonymized data, which can limit the detail and value of their analyses.

However, there are several limitations to the CCLDs. First, hospitals having the discretion to select which data items to include led to variations in the types and amounts of data collected. In the real world, where each institution uses a different healthcare information system, differences in collectable data are inevitable. To minimize this disparity, the Cancer Data Standards and Quality Committee provided guidance, suggesting that hospitals gather data items that can be extracted in a structured format. Furthermore, all hospitals are required to achieve a collection rate of at least 60% and strive to improve their collection rate each year. Metadata information on the tables and columns collected by each institution is now available and will be updated annually on the K-CURE portal. Second, multicenter research can be hindered by a complex approval process. Researchers must obtain approval from the IRB and DRB at every participating hospital. However, this is the first national initiative that provides researchers with access to clinical data from hospitals. The Korean government continues to work toward simplifying the approval process while ensuring legal compliance.

In conclusion, in this article, we introduce the CCLD, a pioneering standardized cancer-specific database that overcomes the limitations of previous CDWs. The use of clinical
data has traditionally encountered challenges pertaining to limited accessibility due to privacy concerns and issues with unstructured data across multiple sites. We have addressed these challenges by creating standardized cancer-specific CCLDs containing de-identified information. These CCLDs are exclusively assessable via a secure environment. The CCLDs are Korea’s largest-scale clinical cancer databases and are the result of a collaborative effort by experts in oncology, healthcare, and data science from leading hospitals across South Korea. We aim to have CCLDs established for ten primary cancers by 2025. Their successful development holds significant promise as a valuable resource for multicenter cancer research.
Author Contributions

Conceived and designed the analysis: Choi KS, Kong HJ, Cha H, Kim HJ, Ryu KS, Jeon YS, Kim H, Jung J, Im JS, Chae H.


Wrote the paper: Lee S, Chae H.

ORCID iDs

Sangwon Lee: https://orcid.org/0000-0002-2587-8915

Heejung Chae: https://orcid.org/0000-0003-3547-3521

Conflicts of Interest

Conflict of interest relevant to this article was not reported.

Acknowledgments

We would like to express our gratitude to the Korean Ministry of Health and Welfare and the Korea Health Information Service for their support of the K-CURE project. We extend our thanks to the fifteen participating hospitals in Korea (National Cancer Center Korea, Asan Medical Center, Samsung Medical Center, Seoul National University Hospital, Severance
Hospital, Korea University Anam Hospital, Pusan National University Hospital, Seoul National University Bundang Hospital, Ajou University Hospital, Gachon University Gil Medical Center, Konyang University Hospital, Hallym University Sacred Heart Hospital, Jeonbuk National University Hospital, Chonnam National University Hwasun Hospital, and Daegu Catholic University Medical Center) for their substantial involvement in the development of the CCLD. We also acknowledge the valuable contributions of academic associations (Korean Breast Cancer Society, Korean Gastric Cancer Association, Korean Liver Cancer Association, Korean Society of Coloproctology, Korean Association for Lung Cancer, Korean Association for Thoracic Surgical Oncology, Korean Prostate Society, Korean Urological Oncology Society, Association of Hepato-Biliary-Pancreatic Surgery, Korean Pancreatobiliary Association, Korean Society of Pediatric Hematology-Oncology, Korean Society of Gynecologic Oncology, Korean Society of Medical Oncology, Korean Society for Radiation Oncology) in defining the “Standard Item Definition Guidelines for Cancer.” This paper is supported by the Ministry of Health and Welfare (No. 22A2400-1)
Reference


11. Kim J, Son JH, Kong TW, Chang SJ. Gynecologic cancer clinical practice guidelines in

Table 1. Description of data items collected in Cancer Clinical Library Databases (CCLDs)

<table>
<thead>
<tr>
<th>Class</th>
<th>Contents</th>
</tr>
</thead>
</table>
| Patient health information | Cancer characteristics and patient demographic collected from the National Cancer Registration Program, medical history, alcohol and smoking habit, family history.  
- Items specific to breast cancer: obstetric history. |
| Clinical assessment        | Physical measurement, performance status, chief complaint, diagnosis, clinical staging, metastasis                                         |
| Test result                | Laboratory test, imaging test, biopsy and surgical pathology  
- Items specific to breast cancer: Germline genetic test, gene expression test  
- Items specific to gastric cancer: H. pylori test, esophagogastroscopy, endoscopic ultrasound |
| Treatment                  | Surgery, radiation therapy, systemic therapy, other medication, transfusion, treatment-related complication  
- Items specific to gastric cancer: endoscopic resection |
| Follow-up observation      | Hospital visit, recurrence, death                                                                                                      |
**Fig. 1.** Hospitals participating in the Korea-Clinical Data Utilization Network for Research Excellence (K-CURE) project. Fifteen hospitals across different regions in Korea are in the process of establishing Cancer Clinical Library Databases (CCLDs) as a part of the K-CURE project.
**Fig. 2.** Overview of the two-step CCLD data item standardization process. In the first step, items were listed based on their importance for cancer research in collaboration with academic societies. In the second step, the final items have been selected considering the feasibility of collection by hospitals, followed by the creation of a standard schema for the CCLDs by the Cancer Data Standards and Quality Committee.
Fig. 3. Storage of patient clinical information. Patient clinical data is collected and stored in the respective tables throughout the entire diagnosis, treatment, and follow-up process.
Fig. 4. Overview of cancer type-specific hospital Cancer Clinical Library (CCLD) establishment process. Hospitals extract cancer patient data from their built-in clinical data warehouses and store the information in a cancer type-specific CCLD.
Fig. 5. Initial collection and update of data for CCLDs. Initial collection and update of data for CCLDs. A CCLD initially collects patient clinical data from 2010 to the year before its establishment. Following its establishment, the CCLD is updated to include. Subsequent follow-up data as well as data for newly registered patients on an annual basis.
**Fig. 6.** Procedure for using Cancer Clinical Library Databases (CCLDs) for multicenter cancer research. Researchers are allowed to access CCLD data after obtaining Institutional Review Board and Data Review Board approval. Analyses on the data may only be performed in a secured environment.