Advancing Korean nationwide registry for hepatocellular carcinoma: a systematic sampling approach utilizing the Korea Central Cancer Registry database

Bo Hyun Kim1,7*, E Hwa Yun2*, Jeong-Hoon Lee3,7, Geun Hong4,7, Jun Yong Park5,7, Ju Hyun Shim6,7, Eunyang Kim2, Hyun-Joo Kong2, Kyu-Won Jung2, Young-Suk Lim6,7

1Division of Gastroenterology, Center for Liver and Pancreatobiliary Cancer, National Cancer Center, Goyang; 2Division of Cancer Registration and Surveillance, National Cancer Control Institute, National Cancer Center, Goyang; 3Department of Internal Medicine and Liver Research Institute, Seoul National University College of Medicine, Seoul; 4Department of Surgery, Ewha Womans University College of Medicine, Seoul; 5Department of Internal Medicine, Yonsei University College of Medicine, Seoul; 6Department of Gastroenterology, Asan Medical Center, University of Ulsan College of Medicine, Seoul; 7Korean Liver Cancer Association, Seoul, Korea

Hepatocellular carcinoma (HCC) presents a substantial public health challenge in South Korea as evidenced by 10,565 new cases annually (incidence rate of 30 per 100,000 individuals), in 2020. Cancer registries play a crucial role in gathering data on incidence, disease attributes, etiology, treatment modalities, outcomes, and informing health policies. The effectiveness of a registry depends on the completeness and accuracy of data. Established in 1999 by the Ministry of Health and Welfare, the Korea Central Cancer Registry (KCCR) is a comprehensive, legally mandated, nationwide registry that captures nearly all incidence and survival data for major cancers, including HCC, in Korea. However, detailed information on cancer staging, specific characteristics, and treatments is lacking. To address this gap, the KCCR, in partnership with the Korean Liver Cancer Association (KLCA), has implemented a systematic approach to collect detailed data on HCC since 2010. This involved random sampling of 10-15% of all new HCC cases diagnosed since 2003. The registry process encompassed four stages: random case selection, meticulous data extraction by trained personnel, expert validation, anonymization of personal data, and data dissemination for research purposes. This random sampling strategy mitigates the biases associated with voluntary reporting and aligns with stringent privacy regulations. This innovative approach positions the KCCR and KLCA as foundations for advancing cancer control and shaping health policies in South Korea.

Keywords: Carcinoma, hepatocellular; Registries; Epidemiology; Incidence; Neoplasm staging

INTRODUCTION

Hepatocellular carcinoma (HCC) is the sixth most common cancer and ranks as the second most common cause of cancer-related deaths in South Korea.1 To improve patient care, it is crucial to understand cancer trends and outcomes. A cancer registry is a data system designed to collect, store, and handle information on individuals diagnosed with cancer. Collecting accurate, complete, and unbiased information on cancer is important for effective cancer control, epidemiological research, and planning of public health initiatives.2,3 These activities would be the cornerstone for mitigating the burden of cancer.

The Korea Central Cancer Registry (KCCR), legislated by the Cancer Control Act, is a comprehensive cancer registry institution...
established by the Ministry of Health and Welfare in 1980. The KCCR constructed the Korea national cancer incidence database by merging ad-hoc medical review surveys and national mortality data from statistics Korea, which was initiated as a hospital-based registry but transitioned to a population-based registry in 1999. The registry encompasses over 95% of new cancer diagnoses in Korea, with contributions from hospitals across the country. The KCCR releases the annual report of cancer statistics in Korea each year; however, the registry’s information regarding detailed cancer characteristics or treatment modalities is somewhat limited.\(^1\,4\,5\)

Cancer control planning and evaluation rely on data to inform planning, resource allocation, and progress assessment.\(^6\) Investigating all detailed information on cancer patients, beyond just incidence and mortality data, would provide the most comprehensive insights; however, this is often impractical because of logistical, ethical, and financial constraints. Substantial national cancer registries, such as the KCCR, collect data on cancer incidence and mortality, which may or may not include summary staging information. Summary staging categories are commonly applied to various solid cancers and are broad and simple. Therefore, it is difficult to analyze the specific characteristics of each cancer type. In contrast, other cancer registries are usually based on voluntary registration by researchers. While researcher-initiated registries can collect specific data, they may also be susceptible to selection bias.

To bridge this gap, it is imperative to employ an unbiased sampling strategy as the foundation of the study design. The random selection of a sample that accurately represents the broader population is a critical methodological step to ensure that the findings are generalizable and not skewed by selection bias. Ideally, this sample should have been drawn from the entire national population. This randomization process mitigates the influence of confounding variables and strengthens the reliability of the analyses and conclusions, thereby providing a guide for effective public health interventions and policy decisions.

**HISTORY**

**Voluntary registration method**

The Korean Liver Cancer Association (KLCA), formerly known as the Korean Liver Cancer Study Group, initially formed a Committee for Primary Liver Cancer Registry in 2003 and began a web-based database for the Primary Liver Cancer Registry in 2004 to collect representative and detailed information on primary liver cancer in Korea.\(^7\)

Members accredited by the KLCA can enter data on their own patients with primary liver cancer (mostly HCC) that they manage and select. The KCCR annually verifies and updates the survival information of patients in the web-based registry database, thereby tracking each patient’s survival outcomes. The registry database includes mandatory fields, such as personal identification number, diagnosis method, date of HCC diagnosis, and date of the last confirmation of life or death as well as fundamental items, such as tumor stage, Child-Pugh classification, underlying diseases, medical history, and treatment modalities.

**Two-stage sampling method**

The KLCA has pursued enhancing its data representativeness and quality because regional imbalances and variances in survival data between voluntary registry participants and the broader HCC cohort were noted, and the selected patient data did not represent clinical manifestations and practice in the real world. In 2010, the KLCA performed a random sampling registration project in collaboration with the KCCR to reflect general HCC demographics more accurately. The KLCA and KCCR planned to enroll 15% of all patients with HCC, taking into account 10% with missing data, and thus randomly sampled and investigated 16.5% of patients diagnosed between 2003 and 2005. Consequently, the demographics, diagnostic methods, underlying etiologies, tumor stages, initial treatment modalities, and survival outcomes of 4,520 patients from 32 hospitals were reported.\(^8\)

With the enactment of the Personal Information Protection Act in September 2011, which mandated explicit consent for health data management, academic institutions faced restrictions on accessing personal identification numbers, which are crucial for verifying patient survival status. However, under the Cancer Control Act, the KCCR can access personal identification numbers. The Korean HCC Registry has been part of the National Cancer Registration Statistics Program since 2013. The KCCR sampled approximately 13% of HCC cases in an in-depth annual survey. Two-stage sampling using a systematic sampling method was adopted instead of a complete random sampling method because of resource limitations for data collection, such as manpower and budget. The Korean HCC Registry has collected records of approximately 19,000 patients with HCC diagnosed between 2008 and 2019 as of December 2023, including information on demographics, diagnostic methods, underlying etiologies, tumor stages, initial treatment modalities, and survival outcomes.
DATA COLLECTION

The KLCA currently administers the Primary Liver Cancer Registry by using a two-stage sampling method in collaboration with the KCCR. This process comprises four essential stages. First, the KCCR conducts two-stage sampling using a systematic sampling method to select cases for inclusion in the registry. Second, trained personnel meticulously extract primary data from medical records. Third, medical doctors affiliated with the KLCA perform a thorough validation of the collected data to ensure accuracy and completeness. Finally, the validated data are distributed to the selected researchers through a designated public application (Fig. 1).

Case definition

The KCCR collects annual data on newly diagnosed cancer cases during the past year from hospitals nationwide. This annual dataset is enhanced by supplementary information from the central and 11 regional cancer registries, which capture cases not recorded in hospital registrations. The compilation and computation of the Korea national cancer incidence database and related cancer statistics were finalized over 2 years. All cancer cases were registered according to the international classification of diseases for oncology, 3rd edition (ICD-O-3) and converted to the international classification of diseases, 10th edition (ICD-10).

After finalizing the Korea national cancer incidence database for 2 years, the KCCR extracted approximately 13% of patients newly diagnosed with HCC (topography codes, C22.0; morphology codes, 8000-8157, 8162-8175, 8190-9136, 9141-9582, 9700-9701) for the corresponding year, based on two-stage sampling using a systematic sampling method. First, the top 50 hospitals, which account for the top 75% of HCC patient registrations nationwide and have available staging information, were selected. Hospitals with a large number of patients have a higher chance of being selected for this step. In the next stage, the patients were sampled within each hospital using a systematic sampling method. The sampling rate was adjusted according to the number of patients at each hospital. Between 2008 and 2019, there was an annual registration of approximately 1,500 patients out of approximately 12,000 incident cases of HCC.9-12

Data abstraction and validation

Next, the KCCR sent a list of sampled patients to the corresponding hospitals. At these institutions, health information managers trained by the KCCR extracted predefined data from medical records for another 1 year. Any data gaps or ambiguities that the managers could not resolve were reported to the KCCR, who then communicated these issues to the Committee for Primary Liver Cancer Registry of the KLCA. In cases where uncertainties arise, consultation with physicians is essential to ensure accuracy and completeness.

Upon receipt of incomplete or ambiguous data notifications from the KCCR, the Committee for Primary Liver Cancer Registry asked the relevant hospital doctors to verify and complete the...
information. Medical doctors, who are members of the KLCA, are responsible for validating missing or uncertain data. After the validation process, the committee returned the refined data to the KCCR, at which point the registry data are completed. The KCCR undergoes data cleaning and links survival status based on death certificate information obtained from statistics Korea, which is one of the most essential pieces of information for the cancer registry. Additionally, and implements de-identification measures to ensure the anonymity and privacy of individuals within its dataset. The vital status of patients in the registry is updated annually.

**Case record forms**

The collected data also included patient clinical characteristics such as age, sex, diagnosis date, serologic markers for viral hepatitis, history of alcohol intake, presence of ascites and encephalopathy, performance status, serum levels of albumin and bilirubin, international normalized ratio, tumor characteristics (such as maximal size and number), intrahepatic structural invasion (portal vein, hepatic vein, hepatic artery, or bile duct), and distant metastasis (Supplementary Table 1). The initial and secondary treatment options include surgical resection, transplantation, transarterial therapy, chemotherapy, radiotherapy, and the best supportive care. Alternatively, a description of the treatment could be provided. A past medical history of antiviral treatment was added to patients diagnosed from 2013 onwards.

**Data distribution**

Every year, the KLCA calls for research proposals using the registry data. KLCA members are eligible to submit applications. Subsequently, the Committee for Primary Liver Cancer Registry reviews and selects approximately 5-10 proposals, and the chosen researchers are granted access to de-identified data for their studies.

**DISCUSSION**

Although some liver cancer registries contain detailed information, many rely on voluntary registration by researchers. Liver cancer registries based on voluntary registration offer the advantage of obtaining further detailed information according to the researcher's intent; however, they may suffer from the limitation of potentially confounding variables, such as the researcher's specialty or treatment outcomes, which can mitigate the representativeness of the entire liver cancer population. Moreover, it mandates that researchers exert additional effort in registering patients, ensuring that the collection of personal health data aligns with the rigorous consent and privacy requirements.

In contrast, liver cancer registries using random sampling are more effective in providing a representative sample of the entire dataset; however, they sometimes encounter challenges in expanding the number of variables without requiring the allocation of human resources for registration. As mentioned previously, the enactment of South Korea's Personal Information Protection Act has necessitated researchers to find a new methodology and switch from a voluntary registry to this two-stage sampling registry using a systematic sampling method because of its stringent data protection criteria. The Primary Liver Cancer Registry in Korea does not require researchers to obtain additional consent for registration because the KCCR can collect incidence and mortality data of cancer patients under the Cancer Control Act and undergo an anonymization process prior to distributing the data to researchers, thereby protecting the privacy of patient information in accordance with legal requirements. To the best of our knowledge, the Primary Liver Cancer Registry in Korea, which is maintained using a systematic sampling method based on national incidence data, is unprecedented. The KLCA published reports in the *Journal of Liver Cancer*, the official journal of the KLCA, on patients diagnosed with HCC between 2008 and 2011, 2012 and 2014, and 2015, respectively.

There are several aspects of this new two-stage sampling method that require improvement. First, using a systematic sampling method requires additional trained personnel to extract the primary predefined data from medical records. To improve the quality of data, it is crucial to educate health data managers periodically. Second, when a list of sampled patients is sent to the hospital where they were diagnosed, there may be inaccuracies if they received treatment at different hospitals because an in-depth survey relies on medical records from diagnosing hospitals. Occasionally, medical data are incomplete or unavailable if patients visit a clinic for a second opinion. A systematic sampling approach was used to exclude these patients from the study. To account for these cases, an additional 10% of the patients were sampled. Another method to solve this issue involves linking treatment information using another source of nationwide health information, such as claims data. However, practical difficulties remain in actual applications.

Understanding the status of patients with cancer is a fundamental step in research, public health strategies, and enhancing patient care. A cancer registry that employs random sampling can provide unbiased and detailed information on cancer patients. Such comprehensive registry data, representative of the entire population, are crucial for the development of public health policies, advancement
Acknowledgments
We express our sincere gratitude to Jae Seok Hwang, Kyung Sup Song, and Hee-Jung Wang for their foundation and dedication to the Korean Primary Liver Cancer Registry.

Conflicts of Interest
None disclosed conflicts of interest on this paper.

Ethics Statement
This review article is fully based on articles which have already been published and did not involve additional patient participants. Therefore, IRB approval is not necessary.

Funding Statement
This work was supported by the National Cancer Center (NCC-2010162), Korea. The funding source had no role in the study design, data curation, or data analysis and interpretation.

Data Availability
Not applicable.

ORCID
Bo Hyun Kim https://orcid.org/0000-0001-8355-3126
E Hwa Yun https://orcid.org/0000-0002-8325-1854
Jeong-Hoon Lee https://orcid.org/0000-0002-0315-2080
Geun Hong https://orcid.org/0000-0003-2597-1068
Jun Yong Park https://orcid.org/0000-0001-6324-2224
Ju Hyun Shim https://orcid.org/0000-0002-7336-1371
Eunyang Kim https://orcid.org/0000-0002-4651-6358
Hyun-Joo Kong https://orcid.org/0000-0003-2319-7195
Kyu-Won Jung https://orcid.org/0000-0002-4389-9701
Young-Suk Lim https://orcid.org/0000-0002-1544-577X

Author contributions
Conceptualization: BHK, YSL
Funding acquisition: BHK
Investigation: BHK, EHY, JHL, GH, JYP, JHS, EK, HJK, KWJ, YSL
Supervision: JHL, GH, JYP, JHS, YSL
Writing - original draft: BHK, EHY, YSL
Writing - review & editing: BHK, EHY, JHL, GH, JYP, JHS, EK, HJK, KWJ, YSL
Approval of final manuscript: all authors

Supplementary Material
Supplementary data can be found with this article online https://doi.org/10.17998/jlc.2024.03.03.

REFERENCES