Data collection framework for electronic medical record–based real-world data to evaluate the effectiveness and safety of cancer drugs: a nationwide real-world study of the Korean Cancer Study Group

Hye Sook Han[†], Kyoung Eun Lee[†], Young Ju Suh[†], Hee-Jung Jee, Bum Jun Kim, Hyeong Su Kim, Keun-Wook Lee, Min-Hee Ryu, Sun Kyung Baek, In Hae Park, Hee Kyung Ahn, Jae Ho Jeong, Min Hwan Kim, Dae Hyung Lee, Siheon Kim, Hyemi Moon, Serim Son, Ji-Hye Byun, Dong Sook Kim, Hyonggin An, Yeon Hee Park and Dae Young Zang ⓘD on behalf of the Korean Cancer Study Group (KCSG)

Abstract

Background: Electronic medical records (EMRs) have the highest value among real-world data (RWD). The aim of the present study was to propose a data collection framework of EMR-based RWD to evaluate the effectiveness and safety of cancer drugs by conducting a nationwide real-world study based on the Korean Cancer Study Group.

Methods: We considered all patients who received ramucirumab plus paclitaxel (RAM/PTX) for gastric cancer and trastuzumab emtansine (T-DM1) for breast cancer at relevant institutions in South Korea. Standard operating procedures for systematic data collection were prospectively developed. Investigator reliability was evaluated using the concordance rate between the recommended input value for representative fictional cases and the input value of each investigator. Reliability of collected data was evaluated twice during the study period at three institutions randomly selected using the concordance rate between the previously collected data and data collected by an independent investigator. The reliability results of the investigators and collected data were used for revision of the electronic data capture system and site training.

Results: Between the starting date of medical insurance coverage and December 2018, a total of 1063 patients at 56 institutions in the RAM/PTX cohort and 824 patients at 60 institutions in the T-DM1 cohort were included. Mean investigator reliability in the RAM/PTX and T-DM1 cohorts was 73.5% and 71.9%, respectively. Mean reliability of collected data in the RAM/PTX and T-DM1 cohort was 90.0% for both cohorts in the first analysis and 89.0% and 84.0% in the second analysis, respectively. Mean missing values of the RAM/PTX and T-DM1 cohorts at the time of simulation of fictional cases and final data analysis decreased from 20.7% to 0.46% and from 18.5% to 0.76%, respectively.

Conclusion: This real-world study provides a framework that ensures relevance and reliability of EMR-based RWD for evaluating the effectiveness and safety of cancer drugs.

Keywords: cancer drug, data collection, medical record, real-world data, relevance, reliability

Correspondence to:
Dae Young Zang
Department of Internal Medicine, Hallym University Sacred Heart Hospital, Hallym University College of Medicine, 22 Gwanpyeong-ro 170 beon-gil Dongan-gu, Anyang-si, Gyeonggi-do 14068, Republic of Korea
**fhdzang@gmail.com,
fhdzang@hallym.or.kr**

Yeon Hee Park
Department of Medicine, Division of Hematology-Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Republic of Korea
yhparkhmo@skku.edu

Hyonggin An
Department of Biostatistics, College of Medicine, Korea University, Anam-dong 5-ga, Seongbuk-gu, Seoul, 02841, Republic of Korea.
hyonggin@korea.ac.kr

Hye Sook Han
Department of Internal Medicine, Chungbuk National University College of Medicine, Chungbuk National University Hospital, Cheongju, Republic of Korea

Kyoung Eun Lee
Department of Hematology and Oncology, Ewha Womans University Hospital, Seoul, Republic of Korea

Young Ju Suh
Department of Biomedical Sciences, College of Medicine, Inha University, Incheon, Republic of Korea

Hee-Jung Jee
Siheon Kim
Hyemi Moon
Serim Son
Department of Biostatistics, College of Medicine, Korea University, Seoul, Republic of Korea

Bum Jun Kim
Department of Internal Medicine, Hallym University Sacred Heart Hospital, Hallym University College of Medicine, Anyang-si, Republic of Korea

Hyeong Su Kim
Department of Internal Medicine, Kangnam Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Republic of Korea

Keun-Wook Lee
Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Republic of Korea

Min-Hee Ryu
Jae Ho Jeong
Department of Oncology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Sun Kyung Baek
Department of Internal Medicine, Kyung Hee University Medical Center, Seoul, Republic of Korea

In Hae Park
Department of Internal Medicine, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Republic of Korea

Hee Kyung Ahn
Department of Internal Medicine, Gachon University Gil Medical Center, Incheon, Republic of Korea

Min Hwan Kim
Department of Internal Medicine, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul

Dae Hyung Lee
Inha University Hospital, Incheon, Republic of Korea

Ji-Hye Byun
Innovation Research Department, Health Insurance Review and Assessment Service, Wonju, Republic of Korea

Dong Sook Kim
Review & Assessment Research Department, Health Insurance Review and Assessment Service, Wonju, Republic of Korea

†These authors contributed equally as first authors.

Introduction

Randomized controlled trials (RCTs) are typically considered the gold standard for generating clinical evidence. Patients enrolled in RCTs are highly selective populations subject to stringent eligibility criteria and managed in strictly controlled settings. This research format facilitates internal validity of the data obtained from RCTs to ensure the highest level of evidence for evaluating the efficacy and safety of new therapies prior to their implementation in clinical practice. However, these features of RCTs may contribute to a lack of external validity and generalizability to real-world clinical practice. It has been estimated that <5% of adult patients with cancer are enrolled in clinical trials. Therefore, it remains unclear whether the findings from RCTs are generalizable to the remaining >95% of the population and whether real-world patients in clinical practice respond to treatment in a manner similar to that in participants in RCTs.[1–3] Therefore, real-world studies using data collected from diverse areas outside the scope of conventional RCTs may improve the generalizability and external validity of the data.[4–6]

In the framework for the United States Food and Drug Administration (FDA)'s Real-World Evidence Program,[7] real-world data (RWD), defined as data related to patient health status and/or the delivery of health care, are routinely obtained from medical sources. Examples of RWD include data derived from electronic health records, pharmacy claims and billing, product and disease registries, and patient-generated data (e.g. patient-reported outcomes and data from mobile or wearable devices). Real-world evidence (RWE) is defined as clinical evidence regarding the usage, risks, and benefits of a medical product derived from RWD analysis.[7] RWD and RWE have gained increased attention as complements to traditional clinical trials. The importance of RWE in oncology is also growing, considering that RCTs do not always account for the entire patient population that are candidates for specific cancer drugs.[2,3] As cancer drug development is evolving rapidly, RWE is required to ensure drug effectiveness and safety.[8–11] Recently, the FDA has accepted RWE to support approvals of original or supplementary indications for oncology drugs. RWE has added value to regulatory submissions for rare or orphan indications as a historical control (avelumab for Merkel cell carcinoma and blinatumomab for acute lymphoblastic leukemia), expanded access data (pembrolizumab for microsatellite instability-high or mismatch repair deficient cancers and lutetium Lu 177 dotatate for gastroenteropancreatic neuroendocrine tumors), and supported supplemental indication approval [palbociclib for male patients with breast cancer (BC)].[12]

Electronic medical records (EMRs) constitute RWD and include demographics, diagnoses, treatment decision-making, prescriptions, and laboratory or imaging data.[13,14] Although EMRs enable the accumulation of large amounts of clinical data and are recognized as definitive data with the highest value among RWD, minimizing the biases that may affect study outcomes remains challenging. Commencement of real-world studies using EMR-based RWD requires study designs that minimize such biases, and data quality management is a key requirement for the extracted data in RWE.

Here, we conducted a nationwide real-world study and proposed a framework for systematically collecting EMR-based RWD to evaluate the effectiveness and safety of cancer drugs. We selected ramucirumab plus paclitaxel (RAM/PTX) and trastuzumab emtansine (T-DM1), which are currently used to treat advanced gastric cancer (GC) and BC, respectively, and systematically collected EMR-based RWD at relevant institutions in South Korea.

Patients and methods

The Korean Cancer Study Group (KCSG) conducted a nationwide real-world study to develop a data collection system for EMR-based RWD for cancer drugs. The consortium of experts for the study comprised oncologists, statisticians, and site management organization. The data collection process for EMR-based RWD is illustrated in Figure 1.

Selection of real-world population

RAM/PTX for GC and T-DM1 for BC have been covered by medical insurance in South Korea since May 2018 and August 2017, respectively. We considered all patients who received each drug between the starting date of medical insurance coverage and December 2018 for enrollment in the study. We excluded patients treated at institutions that included less than two candidate patients, those without institutional
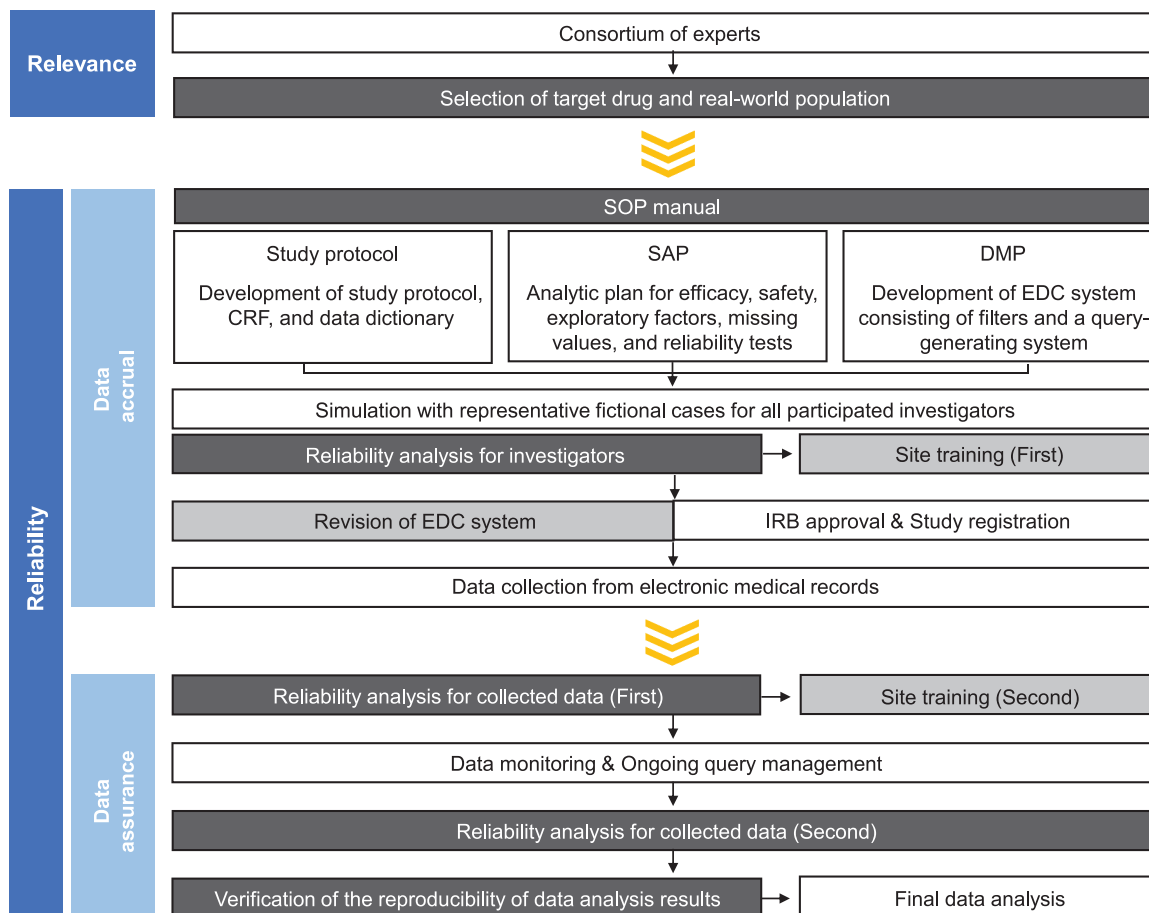
Figure 1. A framework for collection of EMR-based RWD.
CRF, case report form; DMP, data management plan; EDC, electronic data capture; EMR, electronic medical record; IRB, institutional review board; RWD, real-world data; SAP, statistical analysis plan; SOP, standard operating procedure.

review board (IRB) approval, or those who declined to participate. We used eligibility criteria similar to those in the pivotal RAINBOW RCT for RAM/PTX and EMILIA RCT for T-DM1.[15,16] In the RAM/PTX cohort, patients with locally advanced unresectable or metastatic gastric or gastroesophageal junction adenocarcinoma, who experienced disease progression after first-line fluoropyrimidine and platinum-containing combination chemotherapy and received palliative second-line RAM/PTX treatment, were included. In the T-DM1 cohort, patients with human epidermal growth factor receptor 2-positive locally advanced unresectable or metastatic BC, who had previously been treated with trastuzumab and taxane and received palliative T-DM1 treatment were included.

This study was conducted by the GC and BC committees of the KCSG with funding from the Health Insurance Review & Assessment Service

(HIRA). The study was performed in accordance with Good Clinical Practice guidelines and approved by the IRB of each hospital (Supplemental Tables 1 and 2). The IRBs of all participating institutions waived the requirement for informed consent from patients due to the retrospective nature of the study. This study is registered with ClinicalTrials.gov, with identifiers NCT04192734 (KCSG ST19-16) and NCT 04202328 (KCSG BR19-15).

Standard operating procedure manual
To optimize the reliability of the RWD and minimize bias, standard operating procedures (SOPs) are required in retrospective studies using RWD.[7] We prospectively developed an SOP manual that included the study protocol, the common case report form (CRF), a dictionary to define common data elements, a statistical analytic plan (SAP), and a data management plan (DMP) (Supplemental materials).
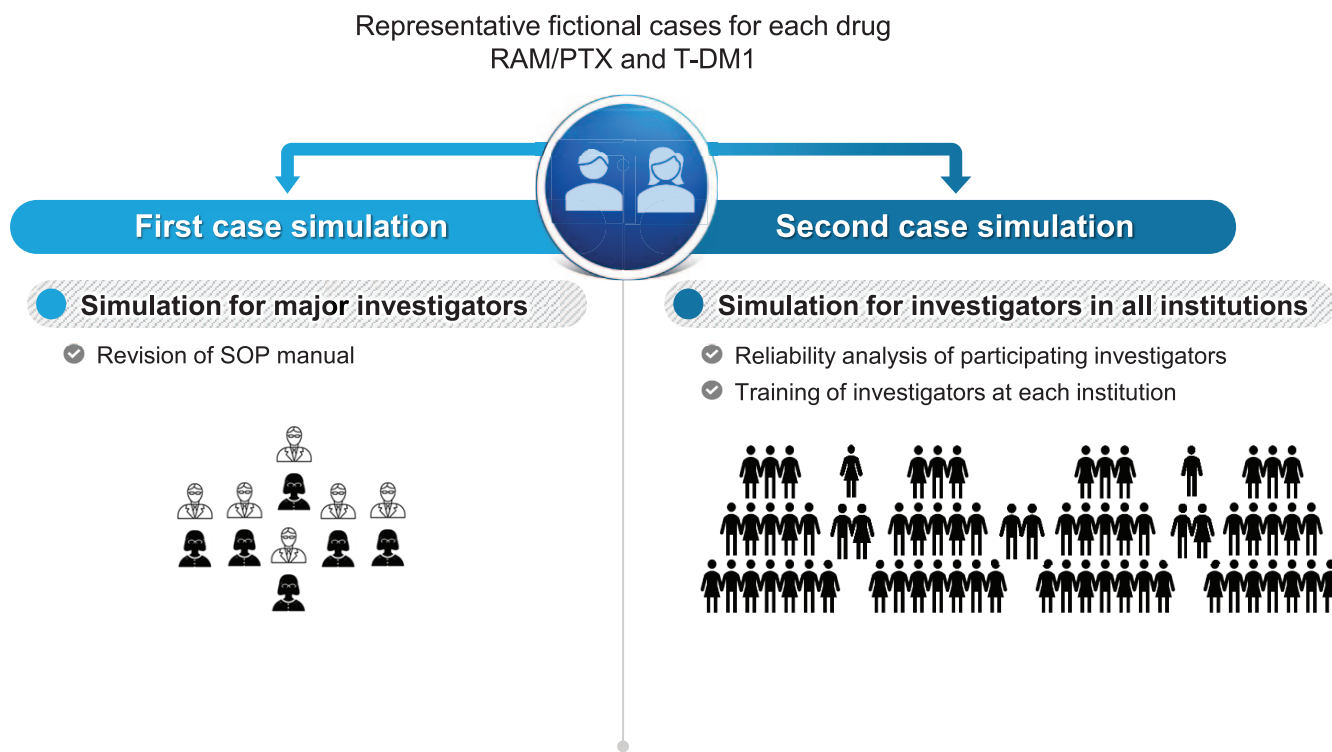
Figure 2. Simulation with representative fictional cases for each drug.
RAM/PTX, ramucirumab plus paclitaxel; SOP, standard operating procedure; T-DM1, trastuzumab emtansine.

Simulation with representative fictional cases and reliability analysis for investigators

We developed two representative fictional cases treated with each drug (Figure 2). The simulation for each case was performed by investigators to analyze the reliability of the participating investigators. Investigator reliability was evaluated using the concordance rate between the recommended input value for representative cases and the input value of each investigator. The inter-investigator agreement was evaluated using the ratio of matching pairs among all possible pairs between investigators. When n investigators evaluate the ith variable with m possible response values (categories), the agreement rate of the ith variable can be obtained as follows:

$$P_i = \frac{1}{n(n-1)}\left(\sum_{j=1}^{m} n_{ij}(n_{ij}-1)\right)$$

where n_{ij} is the number of investigators who responded with category j for the ith variable ($j=1, \ldots, m$).

Data collection and reliability analysis for the collected data

All data were anonymously collected and managed using an electronic data capture (EDC) system consisting of filters and a query-generating system to guarantee reliability and control for errors and missing or inconsistent data. A reliability analysis for collected data was conducted twice during the study period at three institutions randomly selected by statisticians. The first analysis was performed for two institutions at the 50% data collection timepoint. The second analysis was performed for one institution at the completion of data collection. Five patients were randomly selected from each institution and an independent investigator, other than the investigator who collected CRF data from the selected institution, completed the CRF for each case. The reliability of collected data was evaluated using the concordance rate between the previously collected data and data collected by the independent investigator.

Verification of the reproducibility of data analysis results

Two statisticians independently analyzed the same data and compared their results to evaluate

Table 1. Participating institutions and patients.

	RAM/PTX cohort		T-DM1 cohort	
	No. of institutions	No. of patients	No. of institutions	No. of patients
Patients treated with each drug within the study period	92	1460	87	1029
Excluded patients	36	124	27	57
Institutions ⩽ 2 candidate patients	31	48	20	24
Institutions without IRB	1	3	1	3
Institutions that refused to participate	4	73	6	30
Patients assessed for eligibility	57	1336	60	972
Included patients	56	1063	60	824

IRB, institutional review board; RAM/PTX, ramucirumab plus paclitaxel; T-DM1, trastuzumab emtansine.

data analysis reproducibility. The statisticians only shared the study purpose, collected data, and analysis methods or modeling. If the analysis methods or results differed, the most appropriate method and results were determined by consensus with another statistician and oncologists.

Statistical analysis
During CRF development, 'not applicable', which does not require entry, was separately classified and coded from 'missing' values treated as blanks. Missing values were presented as separate data by calculating the missing rate for each variable.

All data are reported as mean and standard deviation (SD) or proportion (%). All analyses were performed using SAS version 9.4 (SAS Institute, Cary, NC, USA).

Results

Real-world population
The participating institutions and patients are shown in Table 1. Between the starting date of medical insurance coverage and December 2018, a total of 1460 patients with GC at 92 hospitals and 1029 patients with BC at 87 hospitals received RAM/PTX and T-DM1, respectively, in South Korea. In the RAM/PTX cohort, 1336 patients at 57 institutions were assessed for eligibility and 1063 eligible patients at 56 institutions were included. In the T-DM1 cohort, 972 patients at 60 institutions were assessed for eligibility and

824 eligible patients were included in the analysis.

Reliability of participant investigators
Reliability of the participating investigators is presented as investigator reliability and inter-investigator agreement rates in Table 2. Reliability of the participant investigators by category for major CRF variables is presented in Figure 3. From the 56 participating institutions in the RAM/PTX cohort, 53 investigators (94.6%) (one per institution) participated in case simulation. The investigator reliability for all 268 variables and 34 major variables of the CRF were $60.5 \pm 27.1\%$ and $73.5 \pm 21.3\%$, respectively (Table 2). The inter-investigator agreement rates including and excluding missing values were $65.0 \pm 21.0\%$ and $86.0 \pm 16.0\%$, respectively (Table 2). CRF variables with investigator reliability and inter-investigator agreement excluding missing values <70% were related to baseline characteristics (age and date of diagnosis for locally advanced or metastatic disease) and adverse events (Figure 3(a) and Supplemental Table 3). CRF variables with a proportion of missing values >10% were associated with RAM/PTX treatment (discontinuation and reasons for discontinuation), adverse events, and survival (date of death or the last follow-up visit) (Figure 3(a) and Supplemental Table 3).

From the 60 participating institutions in the T-DM1 cohort, 58 investigators (96.7%) (one per institution) participated in case simulation. The investigator reliability for all 152 variables and 28 major variables of the CRF were $66.4 \pm 21.6\%$ and

Table 2. Reliability of participating investigators.

RAM/PTX cohort (No. of investigators = 53)		T-DM1 cohort (No. of investigators = 58)	
Investigator reliability (%)[a], mean ± SD			
All CRF variables (n = 268)	Major CRF variables (n = 34)	All CRF variables (n = 152)	Major CRF variables (n = 28)
60.5 ± 27.1	73.5 ± 21.3	66.4 ± 21.6	71.9 ± 16.9
Inter-investigator agreement (%)[b], Mean ± SD			
Including missing values	Excluding missing values	Including missing values	Excluding missing values
65.0 ± 21.0	86.0 ± 16.0	61.0 ± 17.0	83.0 ± 19.0

[a]The concordance rate between the recommended input value for representative fictional cases and the input values of each investigator.
[b]The ratio of matching pairs among all possible pairs between investigators.
CRF, case report form; RAM/PTX, ramucirumab plus paclitaxel; SD, standard deviation; T-DM1, trastuzumab emtansine.
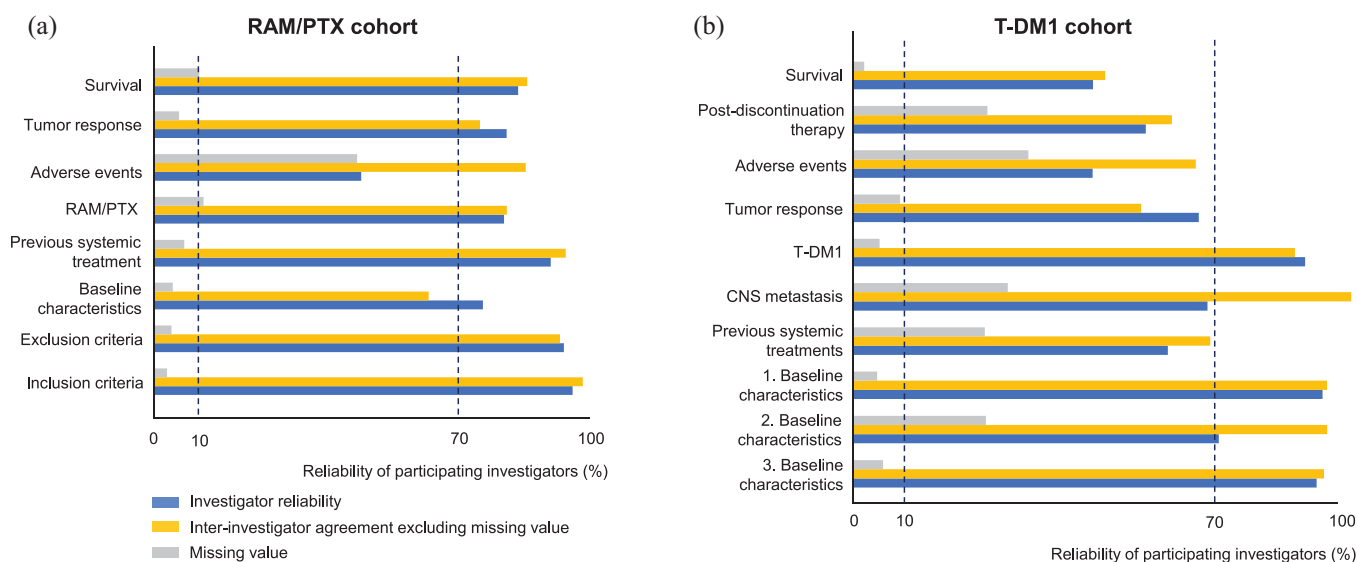

Figure 3. Reliability of the participating investigators and inter-investigator agreement for major variables of case report form in the (a) ramucirumab plus paclitaxel and (b) trastuzumab emtansine cohorts. Each bar represents mean values of variables in each category.
Baseline characteristics of the T-DM1 cohort: 1. baseline characteristics at the diagnosis of locally advanced or metastatic BC; 2. baseline characteristics at initial diagnosis of BC; 3. baseline characteristics at T-DM1 treatment.
BC, breast cancer; CNS, central nervous system; CRF, case report form; RAM/PTX, ramucirumab plus paclitaxel; T-DM1, trastuzumab emtansine.

71.9 ± 16.9%, respectively (Table 2). The inter-investigator agreement rates including and excluding missing values were 61.0 ± 17.0% and 83.0 ± 19.0%, respectively (Table 2). CRF variables with investigator reliability and inter-investigator agreement excluding missing values <70% were associated with baseline characteristics (date of initial diagnosis of BC), previous systemic treatments (protocol), tumor response (date of best response), adverse events, post-discontinuation therapy (protocol), and survival (Figure 3(b) and Supplemental Table 4). CRF variables with a proportion of missing values >10% were related to baseline characteristics at initial diagnosis of BC, previous systemic treatments, central nervous metastasis, adverse events, and post-discontinuation therapy (protocol) (Figure 3(b) and Supplemental Table 4).

Table 3. Reliability of the collected data.

	Reliability of the collected data (%)[a], mean $\pm$ SD			
	RAM/PTX cohort		**T-DM1 cohort**	
	All CRF variables	**Major CRF variables**	**All CRF variables**	**Major CRF variables**
First analysis (No. of patients = 10)	(n = 502)	(n = 33)	(n = 527)	(n = 40)
	90.0 $\pm$ 20.0	90.0 $\pm$ 10.0	90.0 $\pm$ 10.0	90.0 $\pm$ 10.0
Second analysis (No. of patients = 5)	(n = 518)	(n = 32)	(n = 526)	(n = 34)
	88.0 $\pm$ 23.0	89.0 $\pm$ 15.0	94.0 $\pm$ 12.0	84.0 $\pm$ 17.0

[a]The concordance rate between the previously collected data and data collected by independent investigators.
CRF, case report form; RAM/PTX, ramucirumab plus paclitaxel; SD, standard deviation; T-DM1, trastuzumab emtansine.

Reliability of the collected data
For each drug, previously collected data and the data collected by independent investigators were compared for 10 patients from two institutions in the primary analysis and five patients from one institution in the secondary analysis. In the RAM/PTX cohort, the concordance rates for the input values of all CRF variables and major CRF variables in the first analysis were 90.0 $\pm$ 20.0% and 90.0 $\pm$ 10.0%, respectively, and the concordance rates in the second analysis were 88.0 $\pm$ 23.0% and 89.0 $\pm$ 15.0%, respectively (Table 3). In the T-DM1 cohort, the concordance rates for the input values of all CRF variables and major CRF variables in the first analysis were both 90.0 $\pm$ 10.0%, and the concordance rates in the second analysis were 94.0 $\pm$ 12.0% and 84.0 $\pm$ 17.0%, respectively (Table 3). The input data discrepancies between the previously collected data and the data collected by independent investigators were due to the input of variables that were either in textual form or not directly described in the EMRs (performance status, reason for discontinuation, adverse events, best response or disease progression, and survival data), different input date criteria according to each investigator (e.g. variability in examination date, date of result report, date of result confirmation, or date of visit), or incorrect data input based on data collection dates other than the data cutoff date (Supplemental Tables 5 and 6).

Site training and changes in the missing values
First site training adopted the simulation results with representative fictional cases for each drug. Educational information for each institution was delivered by analyzing the variables that had low agreement with the recommended input values of representative cases or high incidence of missing values. EDC system consisting of filtering and a query-generating system were revised to reflect the CRF variables with low reliability of investigators or high incidence of missing values. Subsequent second site training adopted the first reliability results of the collected data conducted at two institutions. Educational information was delivered by analyzing the variables that did not match the input values of investigators at the selected institution and the input values of independent investigators in cross-checking.

Through revision of the EDC system and site training, the missing values of the RAM/PTX and T-DM1 cohorts at case simulation and final data analysis (mean $\pm$ SD) decreased from 20.7 $\pm$ 21.1% to 0.46 $\pm$ 0.59% and from 18.5 $\pm$ 13.7% to 0.76 $\pm$ 1.48%, respectively (Figure 4).

Reproducibility of data analysis results
The analysis results were consistent between statisticians for most variables, but differences were observed in the total number of included patients, time to progression, and progression-free survival. For the included patients, one statistician included all patients with eligibility criteria of 'yes', while the second statistician included patients who met the eligibility criteria and for whom survival data were available (no missing values). For patients with both disease progression and treatment other than the investigational drug in the T-DM1 cohort, one statistician defined the time of disease progression as an event, while the second statistician defined the occurrence of other treatments as censoring. By consensus with a third statistician and oncologist,
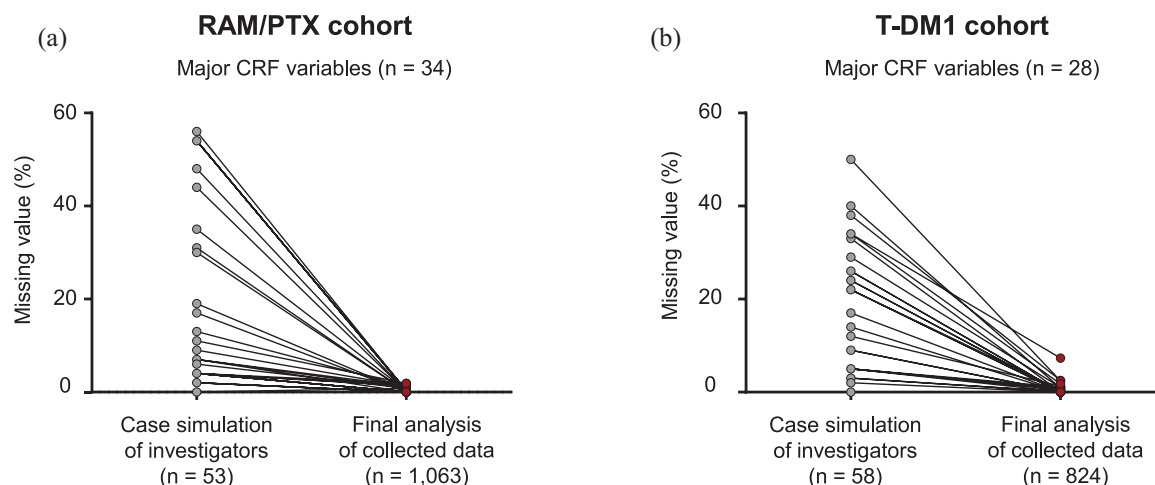
Figure 4. Changes in missing values for major variables of case report in the (a) ramucirumab plus paclitaxel and (b) trastuzumab emtansine cohorts.
CRF, case report form; RAM/PTX, ramucirumab plus paclitaxel; T-DM1, trastuzumab emtansine.

the included patients were defined as those who met the eligibility criteria with available survival data, an event was defined as disease progression or death, and censoring was defined as receiving other treatments or non-occurrence of events. After clarifying the definitions of included patients, events, and censoring, the results regarding the number of included patients, time to progression, and progression-free survival were consistent between the two statisticians.

Discussion
This large-scale, nationwide, real-world study provides a systematic framework for collecting EMR-based RWD with a definitive design and detailed methodologies for data quality management. To the best of our knowledge, no previous study has focused on a detailed data collection process for EMR-based RWD. Previous real-world studies were only based on retrospective data collection and analysis results or were conducted in selected populations.[17–22] Our study is notable in that the systematic collection process for EMR-based RWD was prospectively planned and conducted to minimize bias and increase RWD reliability in nearly all the populations that received candidate drugs.

The KCSG ST19-16 and KCSG BR19-15 studies reported favorable effectiveness and safety of RAM/PTX[23] and T-DM1 and supported the scientific evidence derived from pivotal RCTs. The strengths of such RWE are dependent on the relevance of the underlying data and reliability of data accrual and data assurance (data quality control).[7,24,25] In general, the FDA and other stakeholders assess the relevance and reliability of RWD to determine the suitability of RWD for regulatory decision-making.[7]

The overall assessment of the relevance of RWD determines whether the existing RWD source is appropriate for evaluating drug effectiveness and safety. In the present study, we considered all patients who received RAM/PTX for GC and T-DM1 for BC during the study period and applied similar eligibility criteria to those used in pivotal RCTs that provided evidence for drug approval and reimbursement.[15,16] Therefore, our study employed an all consecutive-comer's design that minimized potential bias in patient selection and enrollment. Furthermore, the real-world population in our study was appropriate for evaluating drug effectiveness and safety.

The reliability of RWD is assessed by evaluating how the data were collected (data accrual) and whether the data collection and analysis procedures provide adequate assurance for error minimization and sufficiency of data quality and integrity (data assurance). In addition, standardization of uniform and systematic methods for collecting and cleaning data is vital for the reliability of RWD. To ensure RWD reliability, SOP for the use of uniform and systematic methods for data

collection and analysis is essential to ensure data quality. The SOP manual in our study described the data elements to be collected, data element definitions, methods for data aggregation and documentation (e.g. CRF form, survey guidelines), SAP (e.g. analytic plan for efficacy and safety, exploratory factors, missing values, and the reliability of investigators and the collected data analyses), and DMP (e.g. EDC system consisting of filters and a query-generating system). In our study, RWD reliability was evaluated based on investigators and the collected data. To evaluate the reliability of investigators, we developed two representative fictional cases treated with each drug. Furthermore, the simulation results with representative fictional cases were used for revision of EDC system and site training. Notably, training all investigators at the participating sites is challenging because real-world studies involve a greater number of institutions and investigators than conventional trials. As such, simulation *via* representative fictional cases is an effective training method that prepares investigators for participation in real-world studies. Data assurance *via* data quality control is essential for providing confidence in the reliability of RWD.[7,24,25] Reliability analysis for the collected data was conducted twice during the study period and evaluated using the concordance rate between previously collected data and data collected by independent investigators. Ongoing additional site training utilized the initial reliability results of the collected data that did not match the input values of investigators in selected institutions and the input values of independent investigators. All the collected data were managed using an EDC system consisting of filters and a query-generating system to control for errors and missing or inconsistent data. Although reliability of the participating investigators was not evaluated again at the time of final data analysis, the reliability of collected data was improved and the missing values were significantly reduced when compared with the extent at the time of case simulation based on the site training and EDC system. Furthermore, to ensure the reproducibility of data analysis results, two statisticians analyzed the same data independently and compared the analysis results. In this regard, RWD assurance can be improved by evaluating data reliability and consistency over time, implementing ongoing training and data monitoring by query-generating systems, and performing independent data analysis by different statisticians.

Despite meticulous data quality management in our data collection process, we identified several variables with low reliability in the second analysis of the collected data. When the data recorded in EMRs were not categorized and were in textual form (e.g. performance status, previous treatments, adverse events, reasons for discontinuation, best response or disease progression, and post-discontinuation therapy), the concordance of investigators was low. In this real-world study, we evaluated the format of the EMRs of patients with cancer receiving RAM/PTX at 55 institutions in South Korea (data not shown). Among CRF variables, approximately 50% of cancer-related or chemotherapy-related variables were described in a textual form in EMRs. The coding of major variables for RWD in the EMRs of cancer patients (e.g. performance status, tumor response, disease progression, and survival) will help to reduce missing data and improve the reliability of EMR-based RWD. Another limitation in the EMR-based RWD collection is that data cannot be collected in the EMRs of each hospital. Even death-related data, which are considered to be the key endpoint in cancer patients, did not exhibit 100% agreement between investigators because EMR data for patients whose death occurred outside the institution were unreliable. In our study, survival could not be accurately determined (indicated as 'unknown' in CRF) in 23.5% patients from the RAM/PTX cohort and 17.7% patients from the T-DM1 cohort. Therefore, it is necessary to link or integrate data with other public institutions that can determine the patient death information while protecting personal information to facilitate reliable RWD collection in cancer patients. Furthermore, there may be human error in our EMR-based RWD collection; big data management systems, such as Clinical Data Warehouse, could not be used because the EMR systems were inconsistent across institutions. Therefore, there is a need for a digital healthcare system herein EMR-based RWD can be structured, defined, formatted, and exchanged with an integrated computer system and converted into scientific data.

The data collection framework for EMR-based RWD proposed in the present study was based on the principles of the FDA's RWE program (Figure 1).[7] The study population included nearly all cancer patients treated with candidate cancer drugs and was representative of routine clinical practice

in South Korea. Furthermore, a detailed SOP manual for collection of uniform RWD and minimization of bias was developed prior to study initiation to ensure data quality. This framework included reliability tests for investigators and the collected data, site training programs, EDC system consisting of filtering and a query-generating system, and verification of the reproducibility of data analysis results. Therefore, the data collection process in our study provides a framework that ensures relevance and reliability of EMR-based RWD for evaluating the effectiveness and safety of cancer drugs. In the future, it will be necessary to develop a big data digital healthcare system that integrates this data collection process and cooperates with public institutions, such as the HIRA and the National Health Insurance Service.

The present framework has several limitations. First, it was inherently limited due to the retrospective observational methodology, including missing data, under-reporting of adverse events and deaths, and the lack of control patients, which could have influenced data analyses for drug effectiveness and safety. Prospective RWD collection would generate evidence based on pragmatic clinical trials that support randomized study designs and extend clinical research to the point of care.[7,26,27] Second, EMR data may not capture all data elements required to answer the question of interest. In that regard, administrative claims may provide a beneficial overlap of information to investigate the balance between costs and benefits.[7,28–32] In addition, novel technologies (e.g. mobile or wearable devices or applications) will enable data collection from candidate patients with cancer in real time. These sources contain vital health-related information, such as demographics, symptoms, long-term effects, adherence rates, and financial burden.[7,33–36] Therefore, future research should focus on developing a data collection framework for RWD derived from prospective studies, pharmacy claims and billing, patient-generated data, and novel technologies. The KCSG has conducted a prospective study (KCSG ST21-06) to establish a RWD collection framework to evaluate not only drug effectiveness and safety, but also quality of life, patient-reported outcomes, financial burdens, and smartphone-based healthcare data (ClinicalTrials.gov with identifiers NCT04915807).

In conclusion, the KCSG has performed a nationwide real-world study and developed a data collection framework to ensure the relevance and reliability of EMR-based RWD for evaluating the effectiveness and safety of cancer drugs. The data collection framework developed here will be a valuable reference for establishing RWE by systematically collecting EMR-based RWD for cancer drugs regardless of the type of cancer and anticancer drugs.

Declarations

Author contribution(s)
Hye Sook Han: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Writing – original draft.

Kyoung Eun Lee: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Writing – original draft.

Young Ju Suh: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Writing – original draft.

Hee-Jung Jee: Data curation; Formal analysis; Investigation; Methodology; Resources; Software.

Bum Jun Kim: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization.

Hyeong Su Kim: Data curation; Investigation; Validation.

Keun-Wook Lee: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Supervision; Validation.

Min-Hee Ryu: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Resources; Software; Supervision; Validation.

Sun Kyung Baek: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation.

In Hae Park: Data curation; Investigation; Methodology; Validation.

Hee Kyung Ahn: Data curation; Formal analysis; Investigation; Methodology; Validation.

Jae Ho Jeong: Data curation; Formal analysis; Investigation; Methodology; Validation.

Min Hwan Kim: Data curation; Formal analysis; Investigation; Methodology; Validation.

Dae Hyung Lee: Formal analysis; Methodology; Validation.

Siheon Kim: Formal analysis; Methodology; Validation.

Hyemi Moon: Formal analysis; Methodology; Validation.

Serim Son: Formal analysis; Methodology; Validation.

Ji-Hye Byun: Conceptualization; Funding acquisition; Methodology; Project administration.

Dong Sook Kim: Conceptualization; Funding acquisition; Methodology; Project administration.

Hyonggin An: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Supervision; Validation; Visualization; Writing – review & editing.

Yeon Hee Park: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Resources; Software; Supervision; Validation; Writing – review & editing.

Dae Young Zang: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Writing – review & editing.

ORCID iD
Dae Young Zang https://orcid.org/0000-0002-2602-7848

Supplemental material
Supplemental material for this article is available online.

References
1. Tannock IF, Amir E, Booth CM, *et al.* Relevance of randomised controlled trials in oncology. *Lancet Oncol* 2016; 17: e560–567.

2. Heng DY, Choueiri TK, Rini BI, *et al.* Outcomes of patients with metastatic renal cell carcinoma that do not meet eligibility criteria for clinical trials. *Ann Oncol* 2014; 25: 149–154.

3. Gan CL, Stukalin I, Meyers DE, *et al.* Outcomes of patients with solid tumour malignancies treated with first-line immuno-oncology agents who do not meet eligibility criteria for clinical trials. *Eur J Cancer* 2021; 151: 115–125.

4. Sherman RE, Anderson SA, Dal Pan GJ, *et al.* Real-world evidence–what is it and what can it tell us? *N Engl J Med* 2016; 375: 2293–2297.

5. Corrigan-Curay J, Sacks L and Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectivene. *JAMA* 2018; 320: 867–868.

6. Rudrapatna VA and Butte AJ. Opportunities and challenges in using real-world data for health care. *J Clin Invest* 2020; 130: 565–574.

7. US Food and Drug Administration. *Framework for FDA's real-world evidence program.* http://www.

fda.gov/media/120060/download (2018, accessed October 2021).

8. Khozin S, Blumenthal GM and Pazdur R. Real-world data for clinical evidence generation in oncology. *J Natl Cancer Inst* 2017; 109: 1–5.

9. Booth CM, Karim S and Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. *Nat Rev Clin Oncol* 2019; 16: 312–325.

10. Karim S and Booth CM. Effectiveness in the absence of efficacy: cautionary tales from real-world evidence. *J Clin Oncol* 2019; 37: 1047–1050.

11. Di Maio M, Perrone F and Conte P. Real-world evidence in oncology: opportunities and limitations. *Oncologist* 2020; 25: e746–e752.

12. Feinberg BA, Gajra A, Zettler ME, *et al.* Use of real-world evidence to support FDA approval of oncology Drugs. *Value Health* 2020; 23: 1358–1365.

13. Coloma PM, Schuemie MJ, Trifirò G, *et al*; EU-ADR consortium. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf* 2011; 20: 1–11.

14. Kim HS, Lee S and Kim JH. Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. *J Korean Med Sci* 2018; 33: e213.

15. Wilke H, Muro K, Van Cutsem E, *et al.* RAINBOW study group. Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): a double-blind, randomised phase 3 trial. *Lancet Oncol* 2014; 15: 1224–1235.

16. Verma S, Miles D, Gianni L, *et al.* EMILIA study group. Trastuzumab emtansine for HER2-positive advanced breast cancer. *N Engl J Med* 2012; 367: 1783–1791.

17. Abdel-Rahman O, Mulder K and Easaw J. Outcomes of ramucirumab plus paclitaxel among patients with previously treated metastatic gastric/lower esophageal cancer: a real-world study. *Am J Clin Oncol* 2021; 44: 158–161.

18. Di Bartolomeo M, Niger M, Tirino G, *et al.* Ramucirumab as second-line therapy in metastatic gastric gancer: real-world data from the RAMoss study. *Target Oncol* 2018; 13: 227–234.

19. Pizzuti L, Krasniqi E, Sperduti I, *et al.* PANHER study: a 20-year treatment outcome analysis from a multicentre observational study of HER2-positive advanced breast cancer patients from the real-world setting. *Ther Adv Med Oncol* 2021; 13: 17588359211059873.

20. Yokoe T, Kurozumi S, Nozawa K, *et al.* Clinical benefit of treatment after trastuzumab emtansine for HER2-positive metastatic breast cancer: a real-world multi-centre cohort study in Japan (WJOG12519B). *Breast Cancer* 2021; 28: 581–591.

21. Vici P, Pizzuti L, Michelotti A, *et al.* A retrospective multicentric observational study of trastuzumab emtansine in HER2 positive metastatic breast cancer: a real-world experience. *Oncotarget* 2017; 8: 56921–56931.

22. Yardley DA, Krop IE, LoRusso PM, *et al.* Trastuzumab emtansine (T-DM1) in patients with HER2-positive metastatic breast cancer previously treated with chemotherapy and 2 or more HER2-targeted agents: results from the T-PAS expanded access study. *Cancer J* 2015; 21: 357–364.

23. Han HS, Kim BJ, Jee HJ, *et al.* Ramucirumab plus paclitaxel as second-line treatment in patients with advanced gastric or gastroesophageal junction adenocarcinoma: a nationwide real-world outcomes in Korea study (KCSG-ST19-16). *Ther Adv Med Oncol* 2021; 13: 17588359211042812.

24. Berger ML, Sox H, Willke RJ, *et al.* Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE special task force on real-world evidence in health care decision making. *Value Health* 2017; 20: 1003–1008.

25. Corrao G and Cantarutti A. Building reliable evidence from real-world data: needs, methods, cautiousness and recommendations. *Pulm Pharmacol Ther* 2018; 53: 61–67.

26. Lund LH, Oldgren J and James S. Registry-based pragmatic trials in heart failure: current experience and future directions. *Curr Heart Fail Rep* 2017; 14: 59–70.

27. Karanatsios B, Prang KH, Verbunt E, *et al.* Defining key design elements of registry-based randomised controlled trials: a scoping review. *Trials* 2020; 21: 552.

28. Lee K, Kim HO, Choi HK, *et al.* Real-world treatment patterns for patients 80 years and older with early lung cancer: a nationwide claims study. *BMC Pulm Med* 2018; 18: 127.

29. Kish JK, Ward MA, Garofalo D, *et al.* Real-world evidence analysis of palbociclib prescribing

patterns for patients with advanced/metastatic breast cancer treated in community oncology practice in the USA one year post approval. *Breast Cancer Res* 2018; 20: 37.

30. Kwon JW, Tchoe HJ, Lee J, *et al.* The impact of national surveillance for liver cancer: results from real-world setting in Korea. *Gut Liver* 2020; 14: 108–116.

31. Wang H, Liu Z, Guo C, *et al.* Health-seeking behavior and barriers to treatment of patients with upper gastrointestinal cancer detected by screening in rural China: real-world evidence from the ESECC trial. *Lancet Reg Health West Pac* 2021; 12: 100181.

32. Morgans AK, Hepp Z, Shah SN, *et al.* Real-world burden of illness and unmet need in locally advanced or metastatic urothelial carcinoma following discontinuation of PD-1/L1 inhibitor therapy: a medicare claims database analysis. *Urol Oncol* 2021; 39: 733.e1–733.e10.

33. Bae WK, Kwon J, Lee HW, *et al.* Feasibility and accessibility of electronic patient-reported outcome measures using a smartphone during routine chemotherapy: a pilot study. *Support Care Cancer* 2018; 26: 3721–3728.

34. Ethier JF, Curcin V, McGilchrist MM, *et al.* eSource for clinical trials: implementation and evaluation of a standards-based approach in a real-world trial. *Int J Med Inform* 2017; 106: 17–24.

35. Cho Y, Zhang H, Harris MR, *et al.* Acceptance and use of home-based electronic symptom self-reporting systems in patients with cancer: systematic review. *J Med Internet Res* 2021; 23: e24638.

36. Bärlund M, Takala L, Tianen L, *et al.* Real-world evidence of implementing eHealth enables fluent symptom-based follow-up of a growing number of patients with breast cancer with the same healthcare resources. *Clin Breast Cancer* 2021; 22: 261–268.

ⓈSAGE journals