Original Paper

The IDEA Framework for Clinician-Led Data Curation in a Hospital Clinical Data Warehouse: Single-Center Implementation Study

Ok Soon Jeong[1], MBA; Hongryull Pyo[2], MD, PhD; Dong Hyun Sinn[3], MD, PhD; Won Chul Cha[4], MD, PhD; Dong Kyung Chang[5], MD, PhD

[1]Data Management Team, Samsung Medical Center, Seoul, Republic of Korea

[2]Department of Radiation Oncology, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Republic of Korea

[3]Division of Gastroenterology, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Republic of Korea

[4]Office of AI Transformation, Department of Digital Health, Department of Emergency Medicine, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Republic of Korea

[5]Department of Digital Health, Department of Internal Medicine, Division of Gastroenterology, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Republic of Korea

Corresponding Author:

Dong Kyung Chang, MD, PhD
Department of Digital Health, Department of Internal Medicine, Division of Gastroenterology
Sungkyunkwan University School of Medicine, Samsung Medical Center
Seoul 06351
Republic of Korea
Phone: 82 1099330266
Email: dkchang@skku.edu

Abstract

Background: Hospital clinical data warehouses (CDWs) contain longitudinal patient-level information distributed across heterogeneous clinical domains and require strict governance, privacy protection, and controlled data access. Although disease-specific registries facilitate structured clinical research, routine hospital CDW workflows often require repeated technical intervention when clinicians iteratively refine cohorts or integrate additional variables. At Samsung Medical Center (Seoul, Republic of Korea), an operational audit revealed a mean turnaround time of 2.7 (SD 1.6) weeks for clinical data extraction requests, highlighting a workflow gap between clinicians' research questions and engineer-mediated database query development.

Objective: This study describes the development and single-center implementation of IDEA (intersection, difference, expansion, attribution), an informatics abstraction framework designed to support authorized clinicians in performing iterative cohort construction and variable integration within a governed hospital CDW.

Methods: IDEA implements 4 user-facing operations: intersection (I), difference (D), expansion (E), and attribution (A). Attribution enables dynamic integration of variables from multiple predefined clinical data domains into an active cohort. IDEA was implemented within the Samsung Medical Center CDW, containing longitudinal electronic health records from 1995 to 2025 and 12 cancer registries. The implementation was evaluated using oncology cross-registry scenarios, descriptive analysis of user actions recorded between January 2024 and December 2025, including 11,460 IDEA operator events, and analysis of 349 user-defined cohorts ("My Registries") established from 2022 to 2025.

Results: IDEA supported interactive cohort construction, refinement, and attribute-based stratification across 12 cancer registries without manual SQL authoring. Clinical end users generated the majority of the 11,460 operator interactions: research nurses accounted for 7918 (69.1%), physicians for 2871 (25.1%), and IT personnel for 619 (5.4%). Attribution was the most frequent operation (8726, 76.1%), suggesting that iterative variable integration into established cohorts was a major use case. Additionally, 349 user-defined cohorts were registered as reusable baseline populations for subsequent cohort curation.

Conclusions: IDEA serves as a complementary pathway for clinician-led data curation within a governed hospital CDW, integrating cohort refinement, variable integration, and reusable cohort management into a single workflow. This approach may support selected routine curation tasks for authorized users working with quality-validated data, while complex requests

continue to require analyst support. Multicenter evaluations are needed to independently assess dataset accuracy, completeness, and institutional transferability.

JMIR Med Inform 2026;14:e99642; doi: 10.2196/99642

Keywords: clinical data warehouses; CDW; data curation; cohort construction; variable attribution; real-world evidence; data consistency; self-service analytics

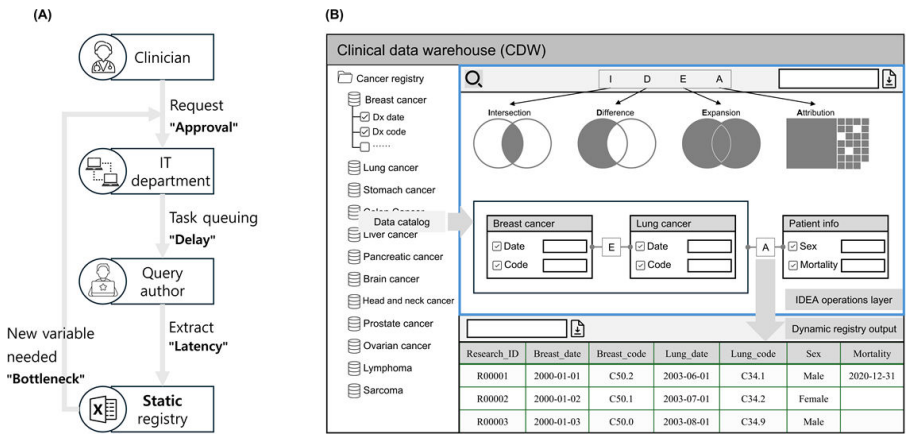
Introduction

Clinical data warehouses (CDWs) have become foundational infrastructure for generating real-world evidence (RWE) and supporting translational health research [1-3]. Hospital CDWs share core data warehousing principles with general-purpose enterprise systems but must additionally accommodate longitudinal patient-level records, heterogeneous clinical domains, mixed structured and unstructured data, evolving medical coding systems, and strict institutional governance and patient privacy requirements [3-6]. Consequently, navigating and extracting clinical data in hospital environments requires clinical context as well as expertise in relational database systems and clinical data management [6]. Furthermore, CDWs face a fundamental challenge in phenotype definition: operational definitions of study cohorts vary substantially not only across investigators but also within the same investigator depending on specific clinical hypotheses. Consequently, it is virtually impossible for centralized IT departments to predefine all potential cohort criteria or build exhaustive disease registries in advance. As clinical hypotheses iteratively evolve, clinicians inevitably require customized cohort restratifications and dynamic variable bindings that static data marts cannot anticipate.

Despite the growing emphasis on the FAIR (findable, accessible, interoperable, and reusable) data principles [7] and advances in cohort discovery platforms and disease-specific data marts [8-13], a workflow gap may remain between the availability of clinical data and clinician-driven hypothesis testing. In conventional hospital CDW workflows, cohort identification, variable integration, and exploratory analysis are often implemented as sequential activities supported by different technical processes. When clinicians iteratively refine inclusion criteria or identify additional variables of interest, each modification may require renewed communication with database analysts to author or modify database-specific SQL logic. Reports from clinical informatics settings have described related challenges involving heterogeneous data sources, data access procedures, governance requirements, and the technical effort required to prepare clinical data for research [4,6,10,14]. The magnitude and operational causes of these challenges may vary across institutions.

This institutional workflow issue was documented at Samsung Medical Center, Seoul, Republic of Korea. In an internal operational audit of 380 consecutive clinical research data extraction requests conducted in 2025, the average turnaround time was 2.7 (SD 1.6) weeks from request submission to final data release. This interval reflected the broader institutional workflow, including clinical requirement clarification, manual query development, and applicable governance procedures, rather than database execution time alone. Although predefined and quality-validated disease registries, including institutional cancer registries, provide structured clinical cohorts [15-18], clinicians at the study site often needed to integrate additional longitudinal attributes from other clinical domains, such as laboratory tests, medications, and clinical notes. Performing these iterative modifications through the conventional request-based workflow required additional technical coordination. Figure 1 illustrates the conceptual distinction between the request-based engineer-mediated workflow and the clinician-led curation workflow evaluated in this study.

Figure 1. Comparison of data curation workflows: (A) Traditional engineer-mediated workflow with operational dependency on IT intermediaries for database-specific query authoring (mean 2.7, SD 1.6 weeks; n=380 consecutive requests) and (B) clinician-led IDEA (intersection, difference, expansion, attribution) framework enabling interactive, end-to-end data curation within approximately 3 minutes through a logic-driven abstraction layer.



In response to this institutional workflow challenge, Samsung Medical Center established a complementary dual-track data architecture rather than attempting to replace centralized informatics and data governance processes. The analyst-mediated engineering pathway (track 1) remains appropriate for computationally complex extractions, institution-wide population cohorts with substantial potential for selection or interpretation bias, and nonstandardized multisystem data integration requiring technical stewardship. The clinician-facing pathway (track 2) supports authorized users working with pre-deidentified, quality-validated clinical data and disease-specific registries within the governed CDW environment [15-19]. The 2 pathways therefore address different operational and governance requirements.

Grounded in these empirical institutional findings and recurring curation requirements, we developed and implemented the IDEA (intersection, difference, expansion, attribution) framework within Samsung Medical Center's governed CDW environment. Rather than introducing a general-purpose theoretical construct independently of the institutional setting, IDEA was developed in response to the specific workflow requirements observed at the study site. The framework organizes iterative clinical data curation into 4 core set-theoretic operators: intersection (I), difference (D), expansion (E), and attribution (A). Attribution enables the dynamic integration of variables from multiple predefined clinical data domains into an active cohort. In addition, evolving cohorts can be saved as user-defined entities ("My Registries") and reused as starting populations for subsequent curation. These functions provide a workflow-oriented abstraction over underlying database operations; they do not imply independence from the underlying relational data structures or query infrastructure.

The objective of this single-center case study was to describe the development, architecture, and institutional implementation of the IDEA framework within a tertiary hospital CDW and to characterize its clinical scenario-based workflow capabilities, user interaction patterns across professional roles, and the longitudinal accumulation of user-defined cohorts registered in the system during routine clinical data curation at the study site.

Methods

Secure Self-Service Architecture and Data Governance

This study involved a retrospective analysis of pseudonymized clinical registry data [19] and system interaction logs. To ensure the protection of personal information [20], data were deidentified at the source level: patient names are removed, and medical record numbers are transformed into nonidentifying codes before integration from the electronic health record (EHR) into the CDW.

To address data governance in a self-service, clinician-led environment, the IDEA framework enforces a dual-layered governance model. First, the logical curation layer supports autonomous logical exploration within a secure research enclave. Because IDEA operates strictly on a pre-deidentified data layer, clinicians can perform continuous cohort refinements and real-time aggregations during preliminary exploration without requiring repeated approvals for individual-level data access during exploratory analyses. Second, the data release layer imposes rigorous access controls: while in-system data manipulation is autonomous, extracting the dynamically curated data matrix outside the secure research environment requires explicit justification and formal approval from the institutional data security committee. This dual-layered approach allows the framework to accelerate hypothesis testing through agile logical curation, while maintaining institutional-level security controls over physical data dispersion.

All data accessed through IDEA are managed within Samsung Medical Center's governed CDW research environment containing pseudonymized clinical data. Access to IDEA requires approval through established institutional procedures [19], and access to disease-specific registries requires separate authorization from the relevant clinical center. To further minimize reidentification risks, potentially identifying information is provided at a restricted level within the research environment (eg, dates of birth are provided only at the month level rather than as exact dates). Clinicians can perform interactive curation within this governed environment; however, export of curated data requires additional security approval, and export of limited-size datasets is restricted by institutional policy (eg, ≤30 records or ≤50 MB).

Importantly, logical cohort refinement occurs entirely within the governed research environment, ensuring that iterative exploration does not require repeated physical movement or duplication of patient-level data.

Architecture and Operational Logic of the IDEA Framework

Overview

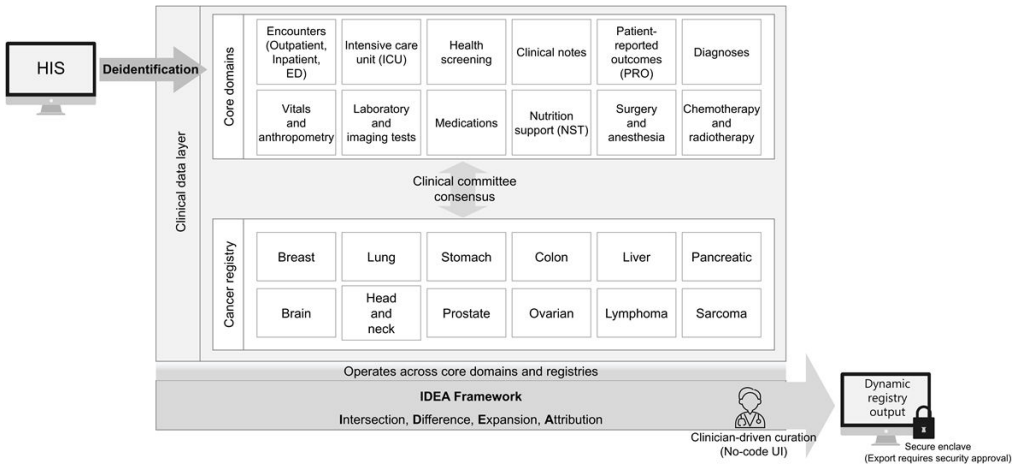
To establish an empirical foundation for the framework, we conducted a qualitative workflow analysis of clinical data extraction requests at our institution. Through this analysis, we identified recurring workflow patterns in how clinicians formulated and refined data requirements. These patterns converged into 2 complementary operational dimensions: horizontal variable expansion and vertical cohort manipulation. Clinicians frequently sought to iteratively bind multidimensional features—such as longitudinal laboratory results, medications, and clinical notes—to an existing patient cohort without restructuring the underlying cohort; this recurring pattern was abstracted as the attribution (A) operator. In parallel, cohort-oriented requests involved isolating overlapping subpopulations that satisfied multiple clinical criteria (intersection [I]), excluding noneligible patient subsets (difference [D]), or combining disease populations across registries (expansion [E]). We also observed that clinicians frequently reused previously derived cohorts as the starting point for subsequent investigations, leading to the recursive "My Registry" mechanism. Thus, rather than being defined a priori, the four IDEA operators and the recursive

"My Registry" mechanism were abstracted from recurring workflow requirements observed in the institutional cases and subsequently formalized within a no-code environment.

This requirement for dynamic abstraction is directly exemplified by real-world institutional studies, such as evaluating data attainability in breast cancer brain metastasis cohorts [18]. In such research, defining an analysis-ready cohort requires composite curation logic: combining breast cancer patients with metastatic central nervous system events (intersection [I] or expansion [E]), excluding non-eligible clinical subsets (difference [D]), and iteratively binding disparate multidomain variables—such as pathology subtypes, systemic therapies, and longitudinal outcome records—to the active patient list (attribution [A]). Analysis of recurring workflow patterns across multiple institutional projects confirmed that clinician-driven data curation fundamentally converges into these 4 atomic operations.

To implement these operations without requiring users to directly author database-specific SQL queries, we designed a layered system architecture. As illustrated in Figure 2, the architecture connects institutional clinical data resources with clinician-driven curation through a layered workflow consisting of a secure clinical data layer, an operational logic layer, and a dynamic registry output layer.

Figure 2. Architecture of the IDEA (intersection, difference, expansion, attribution) framework within the Samsung Medical Center clinical data warehouse. Patient data from the hospital information system (HIS) are deidentified before integration into the clinical data layer, which consists of comprehensive core EHR domains and disease-specific cancer registries established through clinical committee consensus [15-18]. The IDEA framework functions as an informatics abstraction layer operating across both data components to support clinician-led cohort curation within a governed research environment. All logical operations are performed within the deidentified data layer, whereas physical export of generated datasets is subject to institutional security approval. ED: emergency department; EHR: electronic health record.



Data Integration and the Secure Clinical Data Layer

Patient data originating from the Hospital Information System (HIS) are deidentified at source before being integrated into the clinical data layer of the CDW. This layer consists of 2 components: (1) core domains, comprehensive longitudinal EHR elements, including encounters, vital signs, laboratory and imaging tests, medications, clinical notes, and patient-reported outcomes (PROs); (2) disease-specific registries, baseline cohorts (eg, breast, lung, stomach, and colon cancer registries) established through clinical committee consensus and predefined operational criteria.

Certain disease registries, such as the brain cancer registry, capture secondary or metastatic involvement in addition to primary cases. An individual patient's neoplastic disease may also be represented in both site-based and histology-based registries, such as head and neck and sarcoma registries.

Within this governed environment, clinicians construct and refine cohorts using deidentified data. All logical operations are executed within the secure environment; physical export of derived datasets requires separate institutional security approval.

The 4 Operational Operators (I, D, E, A)

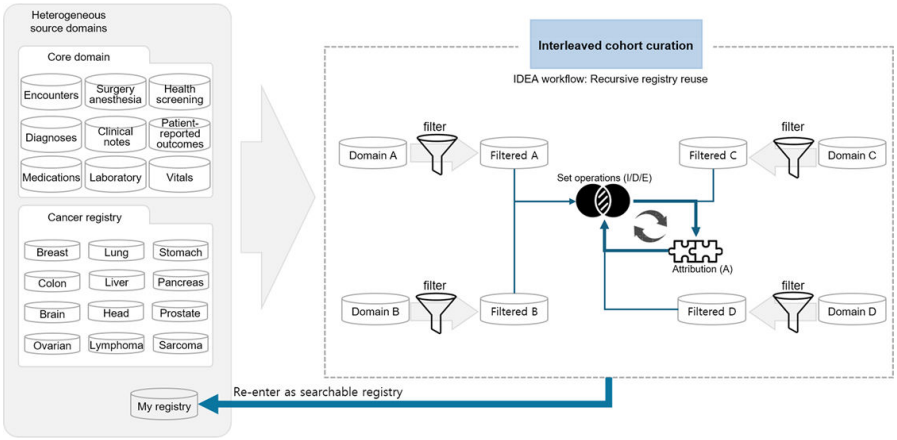
The IDEA framework enables clinician-driven cohort construction through 4 functional operators that abstract complex data manipulation tasks into intuitive clinical operations:

- Intersection (I)–cohort refinement: identifies patients who simultaneously satisfy multiple clinical conditions across different domains, enabling construction of more specific target populations
- Difference (D)–cohort exclusion: removes selected patient subsets from the active cohort to implement clinical exclusion criteria
- Expansion (E)–cohort broadening: combines distinct patient populations across registries or clinical domains to construct broader analytical cohorts
- Attribution (A)–dynamic variable binding: While I, D, and E primarily modify the patient population, the attribution operator links relevant clinical variables to the active cohort. This allows demographic, laboratory, and disease-specific attributes to be added without rebuilding the cohort.

Interleaved Cohort Curation and Recursive Reuse

Figure 3 illustrates the nonlinear workflow of IDEA-based cohort curation. Unlike conventional extraction workflows, in which cohort definition and variable selection are performed as separate steps, IDEA enables clinicians to iteratively alternate between cohort modification and attribute integration within a single interactive environment. Different source domains can be independently filtered, combined using I/D/E operations, and enriched through attribution (A) without losing the current cohort context.

Figure 3. Conceptual overview of the IDEA (intersection, difference, expansion, attribution) framework for interleaved cohort curation and recursive registry reuse. Heterogeneous source domains within the clinical data layer, including core domains and disease-specific registries, are filtered and manipulated through 4 logical operators: intersection (I), difference (D), expansion (E), and attribution (A). Curated cohorts can be saved as user-defined "My Registries" and reintroduced into subsequent workflows as searchable baseline cohorts, enabling iterative refinement and dynamic attribute binding.



Compared with online analytical processing approaches that primarily navigate predefined multidimensional analytical structures, IDEA emphasizes dynamic clinical cohort composition, allowing clinicians to modify the cohort and bind new clinical attributes during curation without rebuilding intermediate datasets.

The resulting dynamic registry output is not treated solely as a final analytical dataset. Instead, curated cohorts can be saved and reintroduced into the framework as user-defined "My Registries," allowing previously defined cohorts to be reused as starting points for further refinement. This recursive mechanism supports the consistent application of cohort definitions across subsequent iterations.

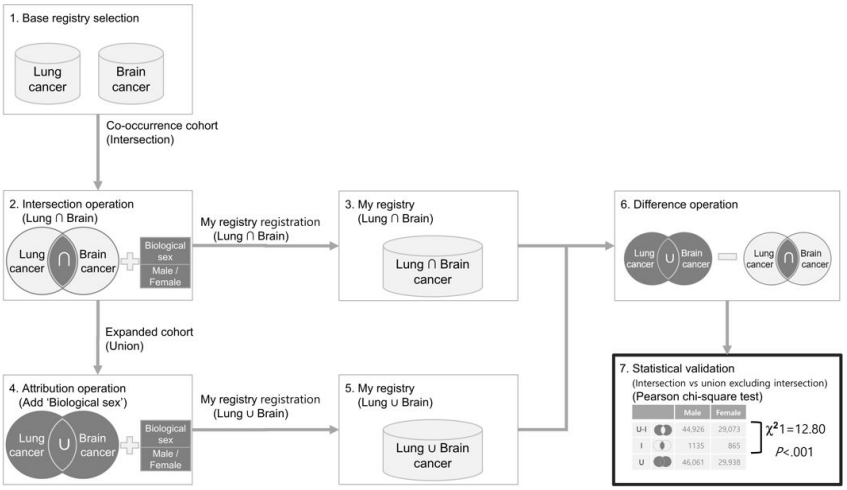
Evaluation Design

The institutional implementation of IDEA was evaluated from three complementary perspectives: (1) workflow expressivity in representative clinical scenarios, (2) real-world usage patterns, and (3) accumulation of user-defined cohorts for reuse.

Clinical Scenario–Based Assessment

To demonstrate the clinical utility and composability of the framework, we evaluated a representative oncology scenario using real-world lung and brain cancer registries (Figure 4 and Multimedia Appendix 1). Clinicians first constructed an intersection (I = L ∩ B) from the lung (L) and brain (B) cancer registries, linked the patient-level attribute of biological sex, and saved the resulting cohort in "My Registry." Subsequently, they generated the corresponding union cohort (U = L ∪ B). To isolate the patients belonging to only one of the 2 registries, a set-difference operation (U – I) was applied to remove the intersecting patients from the union cohort. Finally, the integrated statistical calculator was used to compare the distribution of biological sex between the intersection cohort (I) and the nonoverlapping union subset (U – I) using Pearson χ^2 test. Multimedia Appendix 1 demonstrates this workflow using a Korean-language interface with embedded English subtitles describing each analytical operation and workflow step. All patient identifiers and exact clinical dates displayed in the video were masked using opaque redaction overlays in accordance with institutional data governance policies.

Figure 4. End-to-end multicancer cohort exploration, registry reuse, and statistical validation in the IDEA framework.Workflow steps using lung and brain cancer registries as an institutional use case:Step 1: Selection of baseline disease registries.Steps 2–3: Generation of the co-occurrence cohort (Lung ∩ Brain) using the Intersection (I) operator with "Biological Sex" attribution, registered as a reusable "My Registry."Steps 4–5: Generation of the expanded cohort (Lung ∪ Brain) using the Expansion (E) operator, with "Biological Sex" bound via Attribution (A), and registered as a user-defined cohort.Step 6: Isolation of non-overlapping patients (U – I) using the Difference (D) operator, yielding a cohort independent of the intersection cohort for statistical comparison.Step 7: Pearson's chi-square test comparing sex distributions between the intersection cohort (I) and the union cohort excluding the intersection (U – I) (χ^2=12.80, P.001).All operations are executed within the IDEA environment without manual SQL authoring. Step-by-step user interface operations are provided in Multimedia Appendix 1.



Quantitative Log Analysis

Real-world usage was evaluated using system interaction logs collected between January 2024 and December 2025. Approximately 23,000 user-initiated actions were recorded during the study period. Of these, approximately 11,700 were non-IDEA operations, including single-query lookups and simple noniterative cohort retrievals that did not involve set-theoretic manipulation. The remaining 11,460 actions involved one of the 4 IDEA operators—intersection (I), difference (D), expansion (E), or attribution (A)—and were included in the operator-specific analysis. Each IDEA event was classified according to its operator type and aggregated at the session level. Operator-specific frequencies were summarized descriptively.

Longitudinal Accumulation of Reusable User-Defined Cohorts

In the IDEA framework, a curated registry can be reintroduced as an active input cohort for subsequent refinement or analysis. We quantified the accumulation of user-defined cohorts registered as reusable baseline populations within the system from 2022 to 2025 to characterize the longitudinal establishment of customized registries available for recursive curation.

The evaluation focused on workflow expressivity and observed system usage within the single-center implementation. It did not assess the independent accuracy, completeness, or fitness for purpose of IDEA-generated datasets. Computational latency was not evaluated because execution time may depend on deployment-specific hardware, database configuration, and network conditions.

Ethical Considerations

This study was formally evaluated and determined not to constitute human subjects research by the institutional review board (IRB) of Samsung Medical Center (IRB reference number 2026-04-077). Consequently, formal ethical review and the requirement for informed consent were not applicable, as the research involved only retrospective evaluation of deidentified system interaction logs, aggregated administrative metadata, and deidentified clinical disease registry records for cohort reconstruction via the IDEA framework within a secured institutional environment without accessing identifiable personal health information.

Results

Implementation and Evaluation Overview

The IDEA framework was implemented within the Samsung Medical Center CDW environment, integrating 12 cancer registries with longitudinal clinical data spanning from 1995 to 2025 [15-17]. In this single-center implementation, the framework was evaluated through three complementary approaches: (1) clinical scenario-based assessment to evaluate dynamic cohort generation and cross-registry analysis capabilities, (2) quantitative analysis of real-world user interaction logs to characterize use patterns across professional roles, and (3) quantification of user-defined custom datasets ("My Registries") registered as reusable baseline cohorts for subsequent curation.

These evaluations collectively characterized workflow expressivity, observed system usage, and the longitudinal accumulation of reusable user-defined cohorts within the institutional CDW environment.

Clinical Scenario-Based Assessment for Dynamic Cohort Generation and Cross-Registry Analysis

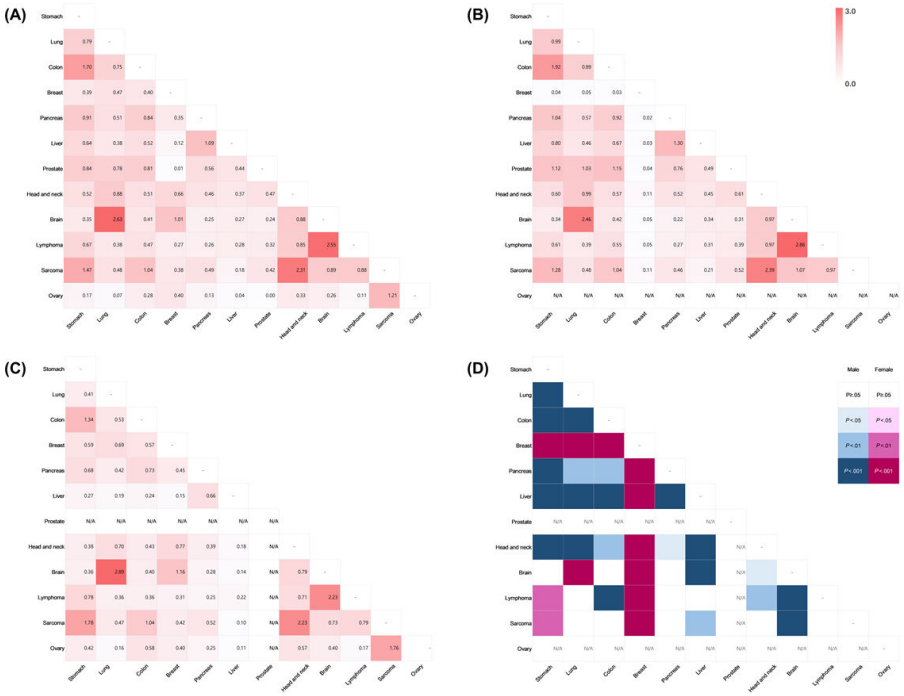
Assessment Overview and Heatmap Interpretation

The clinical scenario-based assessment examined the application of the IDEA framework to representative cross-registry clinical inquiries. By integrating multiple cancer registries with clinical attributes from broader CDW domains, the framework generated analytical datasets for population-level descriptive analysis across heterogeneous clinical data sources. The analyses describe cross-registry cancer co-occurrence patterns and attribute-based stratification obtained through the evaluated curation workflows.

Population-Level Cross-Cancer Co-Diagnosis Mapping

The IDEA framework generated a population-level co-diagnosis map across the 12 cancer registries (Figure 5A-C), illustrating cross-registry patient overlaps across the study population. Several observed patterns reflect known clinical disease trajectories, including secondary or metastatic presentations (such as lung cancer with secondary brain involvement captured within disease registries [21,22]), as well as potential overlapping capture between site-based and histology-based registries, rather than confirmed multiple primary malignancies alone.

Figure 5. Population-level analysis of cancer co-diagnosis patterns and sex-specific disparities enabled by IDEA (intersection, difference, expansion, attribution) operators. Heatmaps illustrate the pairwise co-diagnosis rates (%) among 12 major cancer registries, defined as the Jaccard similarity coefficient (the size of the intersection divided by the size of the union, multiplied by 100) computed through iterative intersection (I) and union (U) operations. (A) Overall population, (B) males, and (C) females. (A-C) Uses a synchronized color scale to facilitate direct visual comparison of prevalence. (D) Sex-specific differences identified after stratifying the cohorts by the biological sex attribute using the attribution (A) operation. Color intensity in panel D represents the statistical significance of sex-based differences (Pearson χ^2 test), with navy indicating a higher relative co-diagnosis rate in males and magenta in females. Cells marked "N/A" denote biologically sex-specific combinations (eg, prostate and ovarian cancer). This mapping, encompassing pairwise comparisons across the 12 cancer registries, illustrates the framework's support for cross-registry cohort composition and attribute-based demographic stratification within a single workflow. Pairwise comparisons were unadjusted for multiple testing.



Sex-Specific Variation in Cancer Co-Diagnosis Patterns

Sex-based stratification of cancer cohorts revealed sex-specific variation in cancer co-diagnosis patterns (Figure 5D). Statistically significant differences between male and female patients were identified using Pearson χ^2 tests for eligible cancer combinations. The heatmap illustrates the magnitude and direction of sex-specific associations across cancer type pairs.

Quantitative Analysis of Real-World User Interaction Logs

User Participation and Operator Events by Professional Role

During the study period (January 2024 to December 2025), approximately 23,000 user-initiated actions were recorded from 46 unique users. Of these, 11,460 were IDEA operator events involving intersection (I), difference (D), expansion (E), or attribution (A); approximately 11,700 were

non-IDEA operations, including single-query lookups and simple noniterative cohort retrievals without set-theoretic manipulation. Among the 11,460 IDEA operator events, research nurses accounted for 7918 (69.1%), physicians for 2871 (25.1%), and IT staff for 619 (5.4%; Table 1). Thus, clinical end users generated most IDEA operator events in this institutional dataset.

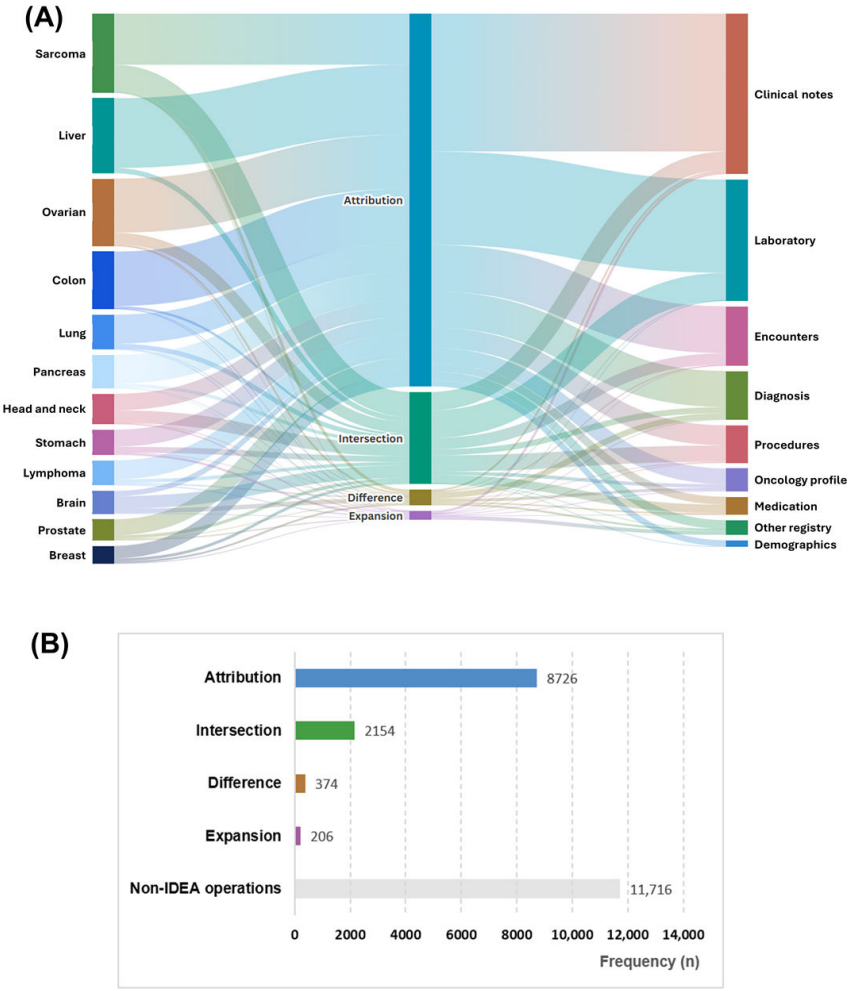
Figure 6 visualizes the observed interaction patterns, including the flow of curation activities and the relative frequencies of the four IDEA operations.

Table 1. Distribution of IDEA (intersection, difference, expansion, attribution) operator events by professional role (total IDEA operator events: 11,460[a]).

Professional role	Unique users (N=46)	Interaction logs (N=11,460), n (%)
Research nurse	15	7918 (69.1)
Medical doctor	12	2871 (25.1)
IT staff	16	619 (5.4)
Research personnel	3	52 (0.5)
Total	46	11,460 (100)

[a]Includes the 11,460 IDEA operator events and excludes approximately 11,700 non-IDEA operations.

Figure 6. Quantitative analysis of observed IDEA (intersection, difference, expansion, attribution) usage patterns and cross-domain data linkage. (A) Alluvial diagram showing the observed flow from source disease registries through IDEA operations to target clinical domains. (B) Bar chart showing the absolute frequencies of IDEA and non-IDEA operations. Non-IDEA operations included single-query lookups and simple noniterative cohort retrievals without set-theoretic manipulation. The 4 IDEA operator events totaled 11,460 and were classified as intersection (I), difference (D), expansion (E), or attribution (A). Attribution accounted for 8726 of 11,460 (76.1%) IDEA operator events, excluding approximately 11,700 non-IDEA operations.



Distribution of IDEA Operator Events

Attribution (A) was the most frequently recorded IDEA operator, accounting for 76.1% (8726/11,460) of IDEA operator events. Intersection (I), difference (D), and expansion (E) were also used to define or modify baseline study populations. These percentages describe the distribution of IDEA operator events at the study site and do not establish the relative difficulty, importance, or causal role of variable integration compared with cohort identification.

Cross-Domain Data Linkage Patterns

In the observed sessions, workflows frequently began with disease-specific registries, including lung, liver, and ovarian cancer registries, followed by attribution operations involving target CDW domains (Figure 6A). Clinical notes and laboratory results were among the target domains with frequent recorded linkages. These findings describe the observed distribution of cross-domain interactions within the study-site log data.

Quantification of User-Defined Cohorts ("My Registries") Established for Reuse

To characterize the establishment of reusable cohorts within the institutional implementation, we analyzed user-defined cohorts registered into the IDEA workflow between 2022 and 2025. Over this 4-year period, 349 user-defined cohorts were established as reusable datasets within the system (82 in 2022, 146 in 2023, 61 in 2024, and 60 in 2025). Previously curated cohorts could be reentered as baseline populations for subsequent query and curation workflows while retaining their cohort definitions and operational states. These observations show that user-defined cohorts were retained within the system and made available for subsequent curation, rather than being limited to single-use data extracts.

Discussion

Principal Findings

This single-center case study described the development and institutional implementation of the IDEA framework in the CDW environment of Samsung Medical Center. IDEA was derived from recurring local requirements for iterative cohort refinement, integration of additional clinical attributes, and reuse of previously curated cohorts. Within this institutional setting, the framework supported clinician-led composition and refinement of clinical cohorts through four operations—intersection (I), difference (D), expansion (E), and attribution (A)—and allowed user-defined cohorts to be retained for subsequent curation. The findings characterize the implementation and observed use of IDEA at the study site rather than establishing a general solution for clinical data curation across institutions.

From an operational and governance perspective, the IDEA framework should be viewed not as a replacement for conventional analyst-mediated informatics workflows but as a complementary component of a dual-track operational model. In routine clinical data operations, traditional engineer-mediated workflows (track 1) remain critical for handling complex data extractions, scenarios requiring analyst oversight, and research inquiries that require data beyond predefined resources. Conversely, the clinician-led IDEA track (track 2) operates over predefined, quality-validated disease registries, including the cancer registries used in this study, and enables authorized researchers to perform iterative cohort curation and recursive registry reuse within established security boundaries. This 2-track operational model accommodates diverse clinical research requirements while preserving institutional data governance and technical safeguards.

Implications for Clinician-Led Data Curation

Existing platforms such as Atlas (Observational Health Data Sciences and Informatics) [9], DEDUCE [10,11], Epic Cosmos [12], and TriNetX [13] provide robust query interfaces [6]; however, their operational focus is predominantly oriented toward population-level cohort identification, whereas iterative attribute incorporation and recursive cohort refinement across diverse local domains often involve secondary data extraction steps. In contrast, IDEA was specifically configured within this institutional implementation to support the iterative integration of additional clinical attributes into previously defined cohorts across local relational clinical databases. A key feature of the evaluated implementation is its operation within a governed, predeidentified CDW environment, which supports exploratory analyses under established institutional security protocols and predefined governance approvals.

In our evaluation, the predominance of attribution (A) operations (76.1%, 8726/11,460) was consistent with an iterative data curation workflow. Although this observation was derived from a single institution and does not establish a universal pattern of clinical data use, it suggests that clinical data exploration frequently extends beyond initial cohort identification to the integration of heterogeneous clinical variables into previously defined or registered cohorts. IDEA addresses this requirement by enabling clinicians to incorporate multidimensional attributes, including clinical documentation and laboratory data, into existing cohorts without requiring technical knowledge of the underlying database structure.

Data Governance and the Evolving Role of Data Engineers

While replacing engineer-mediated workflows with clinician self-service raises valid concerns regarding researcher-introduced bias and data handling errors [14], these risks are heavily mitigated within the IDEA environment through established governance protocols. Data access is not entirely unconstrained; for instance, accessing specific datasets such as cancer registries requires prior authorization, and all data export processes mandate formal security approvals. Furthermore, because cancer and personal registries within the system have already undergone rigorous prior data

validation, using these established datasets actually ensures higher data consistency and reliability.

Consequently, the IDEA framework does not seek to completely eliminate the role of data professionals. Instead, it reallocates their expertise. The intervention of trained data engineers remains essential in specific scenarios, such as supporting novice or infrequent users who may find UI operations burdensome or managing data extraction requests that involve highly complex logic. Finally, the consistency of IDEA-generated analytical patterns with established clinical and epidemiological knowledge (including known secondary involvements such as lung-to-brain metastatic presentations) supports the clinical plausibility of the framework for real-world data exploration within our institutional environment. Although the logic-driven architecture of IDEA suggests conceptual portability across institutional settings, its empirical transferability has not yet been established. Institutions with comparable deidentified CDW environments, standardized clinical data domains, and established governance procedures may be more suitable for future adaptation; however, local implementation and validation studies will be required to assess its applicability in different settings.

Unlike conventional cohort construction pipelines, IDEA introduces a state-preserving and compositional framework for iterative clinical data curation. Existing analytics environments generally separate cohort identification, variable integration, and downstream analysis into sequential processes, requiring repeated extraction and reconstruction when new clinical attributes or hypotheses are introduced. In contrast, IDEA's four operations (intersection, difference, expansion, and attribution) maintain the evolving cohort as an active analytical state within a unified environment, enabling continuous refinement without interrupting the curation workflow.

This architectural characteristic provides a distinct approach to interoperability and multi-center generalizability. While common data models such as OMOP-CDM facilitate interoperability through harmonization of physical data structures [9,23], IDEA focuses on interoperability of curation logic. By representing cohort construction and variable-binding processes as abstract operational logic rather than database-specific queries, IDEA enables the exchange of executable curation strategies across heterogeneous institutional environments.

Such logic-level interoperability establishes the foundation for "Logic Exchange." When participating institutions harmonize core clinical domains (eg, encounters, medications, and diagnoses) through a shared data model, specialized disease registries can be reconstructed by transferring IDEA operational logic rather than exchanging physical datasets or rebuilding institution-specific queries. For example, cancer registry curation rules developed through expert consensus at a tertiary medical center could be disseminated across collaborating institutions, allowing replication of disease-specific registries and facilitating multi-center observational research.

Therefore, IDEA extends interoperability beyond structural data alignment toward reusable knowledge representation, where clinical expertise and curation strategies become transferable computational assets. This approach aligns with prior efforts such as PheKB [23], which promoted the sharing and reuse of computable phenotype definitions, and with FAIR principles [7] (findability, accessibility, interoperability, and reusability) by promoting transparent, machine-interpretable curation logic.

In this study, we evaluated operational feasibility, workflow transformation, and real-world usage of IDEA within a single-center CDW environment. We did not assess intrinsic data accuracy, completeness, or fitness for purpose, as these require separate validation frameworks and were outside the scope of the present study.

Limitations and Future Directions

This study has several limitations. First, as this study was conducted as a single-center case study at a tertiary care hospital, further validation is required to determine whether the findings can be generalized to multicenter settings. Other institutions seeking to adopt this framework may face institutional or technical barriers, particularly concerning diverse data privacy regulations and pseudonymization requirements. However, provided that an institution's data governance policies are strictly enforced at the point of data ingestion into the CDW, researchers can freely generate dynamic data structures within the IDEA framework. Second, while IDEA currently focuses on structured and semi-structured data, the integration of natural language processing (NLP) or large language models (LLMs) is warranted to fully unlock the value of clinical narratives [24,25]. Third, clinician-driven interfaces may introduce potential selection bias. However, this underscores the necessity of domain expertise in informatics; high-performance medicine emerges from the synergy of informatics flexibility and profound clinical knowledge [6,26]. Fourth, it is important to clarify that the reported reduction in data curation time—from an average of 2.7 (SD 1.6) weeks to approximately 3 minutes—represents a contrast between two distinct operational contexts rather than a direct, head-to-head comparative study of equivalent tasks. The 2.7-week figure is derived from an internal audit of 380 real-world requests that include administrative delays and human-to-human communication, whereas the approximately 3-minute duration reflects the technical execution of specific scenarios within the optimized IDEA environment. While these values are not strictly comparable in a controlled experimental sense, the sheer magnitude of the difference underscores a fundamental paradigm shift: moving from an asynchronous, engineer-dependent workflow to a synchronous, self-service clinical informatics environment. Fifth, this evaluation focused on workflow efficiency and system usage within the study site rather than formal data quality assessment. Sixth, cross-registry co-diagnoses generated in this demonstration reflect registry overlap and metastatic disease rather than adjudicating true multiple primary malignancies, requiring subsequent clinical validation when differentiating primary from

metastatic lesions. Future studies should examine whether IDEA-generated datasets are equivalent to independently constructed reference datasets with respect to completeness, correctness, and fitness for purpose.

Conclusions

This single-center case study suggests that the IDEA framework provides a practical, case-derived informatics architecture for supporting clinician-led dynamic data curation within the large-scale CDW environment of Samsung Medical Center. By integrating cohort construction, iterative attribute incorporation, and reusable cohort management into a unified workflow, IDEA enabled interactive exploration of clinical data while reducing reliance on specialized database query authoring for the evaluated scenarios. The predominance of attribution operations (76.1%, 8,726/11,460) and the continued accumulation of reusable user-defined cohorts indicate that within this institutional implementation, users frequently extended existing cohorts through iterative incorporation of additional clinical information. Further multicenter implementation and validation studies will be required to determine the broader applicability and transferability of the framework beyond this local CDW environment.

Generative AI Use

In the preparation of this manuscript, the authors used Gemini (Google) to refine academic expressions and enhance linguistic clarity. Following the use of this tool, the authors rigorously reviewed and revised the content and assume full responsibility for the integrity and accuracy of the final paper.

Data Availability

The datasets analyzed during this study are not publicly available due to privacy, confidentiality, and institutional access restrictions. Requests for access may be directed to the corresponding author and are subject to institutional approval in accordance with the institution's data-governance procedures.

Authors' Contributions

Conceptualization: WCC, DKC
Data curation: HP, DHS
Formal analysis: OSJ
Investigation: HP, DHS
Project administration: DKC
Supervision: DKC
Writing – original draft: OSJ
Writing – review & editing: WCC, DKC

Conflicts of Interest

None declared.

Multimedia Appendix 1

Real-time cohort generation and exploratory analysis via the IDEA framework.
[MP4 File (MP4 video File), 119545 KB-Multimedia Appendix 1]

References

1. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. Nat Rev Clin Oncol. May 2019;16(5):312-325. [doi: 10.1038/s41571-019-0167-7] [Medline: 30700859]
2. 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. Dec 8, 2016;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
3. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. JMIR Form Res. Aug 27, 2020;4(8):e17687. [doi: 10.2196/17687] [Medline: 32852280]
4. Tendedez H, Ferrario MA, McNaney R, Gradinar A. Exploring human-data interaction in clinical decision-making using scenarios: co-design study. JMIR Hum Factors. May 6, 2022;9(2):e32456. [doi: 10.2196/32456] [Medline: 35522463]
5. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. Aug 2013;51(8 Suppl 3):S30-S37. [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
6. Mullin S, Zhao J, Sinha S, Lee R, Song B, Elkin PL. Clinical data warehouse query and learning tool using a human-centered participatory design process. Stud Health Technol Inform. 2018;251:59-62. [Medline: 29968601]
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. Mar 15, 2016;3(1):160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
8. Pop B, Fetica B, Blaga ML, et al. The role of medical registries, potential applications and limitations. Med Pharm Rep. Jan 2019;92(1):7-14. [doi: 10.15386/cjmed-1015] [Medline: 30957080]

9. Hripcsak G, Duke JD, Shah NH, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574-578. [Medline: 26262116]
10. Horvath MM, Rusincovitch SA, Brinson S, Shang HC, Evans S, Ferranti JM. Modular design, application architecture, and usage of a self-service model for enterprise data delivery: the Duke Enterprise Data Unified Content Explorer (DEDUCE). J Biomed Inform. Dec 2014;52:231-242. [doi: 10.1016/j.jbi.2014.07.006] [Medline: 25051403]
11. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. J Biomed Inform. Apr 2011;44(2):266-276. [doi: 10.1016/j.jbi.2010.11.008] [Medline: 21130181]
12. Mankowski MA, Bae S, Strauss AT, et al. Generalizability of kidney transplant data in electronic health records - the Epic Cosmos database vs the Scientific Registry of Transplant Recipients. Am J Transplant. Apr 2025;25(4):744-755. [doi: 10.1016/j.ajt.2024.11.008] [Medline: 39550008]
13. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. JCO Clin Cancer Inform. Dec 2018;2:1-10. [doi: 10.1200/CCI.17.00067] [Medline: 30652541]
14. Kim HH, Kim B, Joo S, Shin SY, Cha HS, Park YR. Why do data users say health care data are difficult to use? A cross-sectional survey study. J Med Internet Res. Aug 6, 2019;21(8):e14126. [doi: 10.2196/14126] [Medline: 31389335]
15. Kim H, Kim HJ, Kim H, et al. Real-world data from a refractory triple-negative breast cancer cohort selected using a clinical data warehouse approach. Cancers (Basel). Nov 21, 2021;13(22):5835. [doi: 10.3390/cancers13225835] [Medline: 34830990]
16. Kim ST, Kang D, Kim SJ, et al. Changes in cancer care for patients aged 80 and above: a cohort study from Samsung Comprehensive Cancer Center in South Korea. Cancers (Basel). Jun 17, 2025;17(12):2017. [doi: 10.3390/cancers17122017] [Medline: 40563667]
17. Jung HA, Jeong O, Chang DK, et al. Real-time automatically updated data warehouse in healthcare (ROOT): an innovative and automated data collection system. Transl Lung Cancer Res. Oct 2021;10(10):3865-3874. [doi: 10.21037/tlcr-21-531] [Medline: 34858777]
18. Kim MJ, Kim HJ, Kang D, et al. Preliminary attainability assessment of real-world data for answering major clinical research questions in breast cancer brain metastasis: framework development and validation study. J Med Internet Res. Mar 23, 2023;25:e43359. [doi: 10.2196/43359] [Medline: 36951923]
19. Shin SY, Lyu Y, Shin Y, et al. Lessons learned from development of de-identification system for biomedical research in a Korean tertiary hospital. Healthc Inform Res. Jun 2013;19(2):102-109. [doi: 10.4258/hir.2013.19.2.102] [Medline: 23882415]
20. Shin SY. Issues and solutions of healthcare data de-identification: the case of South Korea. J Korean Med Sci. 2018;33(5):e41. [doi: 10.3346/jkms.2018.33.e41] [Medline: 29349950]
21. Wan X, Mo H, Chen J, et al. Risk factors for brain metastasis in lung cancer: an umbrella review of systematic reviews and meta-analyses. BMJ Open. Jan 7, 2025;15(1):e087181. [doi: 10.1136/bmjopen-2024-087181] [Medline: 39773793]
22. Schouten LJ, Rutten J, Huveneers HAM, Twijnstra A. Incidence of brain metastases in a cohort of patients with carcinoma of the breast, colon, kidney, and lung and melanoma. Cancer. May 15, 2002;94(10):2698-2705. [doi: 10.1002/cncr.10541] [Medline: 12173339]
23. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. Nov 2016;23(6):1046-1052. [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
24. Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med. Mar 2023;155:106649. [doi: 10.1016/j.compbiomed.2023.106649] [Medline: 36805219]
25. Yan C, Ong HH, Grabowska ME, et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. J Am Med Inform Assoc. Sep 1, 2024;31(9):1994-2001. [doi: 10.1093/jamia/ocae072] [Medline: 38613820]
26. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

Abbreviations

CDW: clinical data warehouse
EHR: electronic health record
FAIR: findability, accessibility, interoperability, reusability
IDEA: intersection, difference, expansion, attribution
IRB: institutional review board
LLM: large language model

NLP: natural language processing
PRO: patient-reported outcomes
RWE: real-world evidence

Edited by Arriel Benis, Michael Leyer; peer-reviewed by Peter Leese, Ross Gore; submitted 28.Apr.2026; initial comments to author 27.Jul.2026; final revised version received 07.Oct.2026; accepted 07.Oct.2026; published 23.Nov.2026

Please cite as:
Jeong OS, Pyo H, Sinn DH, Cha WC, Chang DK
The IDEA Framework for Clinician-Led Data Curation in a Hospital Clinical Data Warehouse: Single-Center Implementation Study
JMIR Med Inform 2026;14:e99642
URL: https://medinform.jmir.org/2026/1/e99642
doi: 10.2196/99642

© Ok Soon Jeong, Hongryull Pyo, Dong Hyun Sinn, Won Chul Cha, Dong Kyung Chang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.