Implementation Report

Development and Evaluation of a Metadata-Driven Rule Repository and Daily Screening System for Diagnosis Coding Compliance Under the Korean New Diagnosis-Related Group Payment System: Implementation Report

Donggyo Shin, PhD; Jinhee Hwang, BS

Department of Medical Information Management, National Health Insurance Service, Ilsan Hospital, Goyang, Gyeonggi-do, Republic of Korea

Corresponding Author:

Donggyo Shin, PhD
Department of Medical Information Management
National Health Insurance Service, Ilsan Hospital
100 Ilsan-ro, Ilsandong-gu
Goyang, Gyeonggi-do 10415
Republic of Korea
Phone: 82 31-900-0800
Email: donggyosss@gmail.com

Abstract

Background: Under the new diagnosis-related group (NDRG) payment system in Korea, the Health Insurance Review and Assessment Service (HIRA) applies disease-specific diagnosis coding approval criteria that differ from the coding guidelines of the Korean Classification of Diseases (KCD), frequently leading to claim denials for conditions such as acute kidney injury, pleural effusion, and malnutrition. Repeated denials also induce conservative down-coding among coders. Because denial notifications are issued 9 to 18 months after discharge, reactive, denial-triggered quality management is insufficient.

Objective: We implemented a metadata-driven rule repository that operationalizes HIRA's review criteria as data, coupled with a daily-refreshed screening data mart (encompassing 6 clinical data domains and 14 data sources) and a robotic process automation (RPA) notification pipeline, to enable proactive, preclaim coding quality management.

Methods: This Implementation Report describes the design, deployment, and 12-month operational experience of the system at an 820-bed NDRG pilot hospital. HIRA's criteria were externalized into a 4-layer rule repository—criteria definitions, clinical concepts, disease-specific attributes, and decision rules. A nightly extract, transform, and load pipeline populates a patient-by-attribute screening data mart within the clinical data warehouse; each morning, Python-based RPA compiles and emails review candidate lists to coders. Descriptive statistics summarized coverage and detection outcomes; precision (equivalent to positive predictive value) was calculated with Wilson score 95% CIs.

Implementation (Results): At the end of the evaluation period, the rule repository comprised 11 review criteria, 271 clinical concepts, 77 attributes, and 96 decision rules. Over the 12-month evaluation period (July 2025 to June 2026), the system evaluated 17,138 diagnosis- or episode-level targets across 15,388 distinct episodes—67% of the hospital's 22,970 NDRG episodes—and flagged 576 (3.4%) review candidates, of which 50 (8.7%, 95% CI 6.6%-11.3%) were confirmed by coders as actual coding errors. The proportion of confirmed coding errors among review candidates differed significantly across the 3 monitored categories (χ^2_2=39.9; P<.001): present-on-admission (POA) coding items showed the highest precision (20/82, 24%), followed by the NDRG cost-outlier category (17/134, 12.7%) and frequently denied diagnoses (13/360, 3.6%). The system was designed, built, and operated by existing departmental staff without additional budget or dedicated IT support, required an estimated 1 to 1.5 hours of coder review time per week, and has continued to operate daily on weekdays since launch.

Conclusions: A metadata-driven rule repository integrated with a daily screening data mart and RPA can operationalize payer review criteria as versioned data rather than hard-coded logic, enabling proactive coding surveillance before claim submission. The approach was implemented without additional budget or dedicated IT staff, but its performance may vary across institutions. Further validation against formal HIRA review outcomes and multi-institutional evaluation are needed to establish its performance and transferability across NDRG hospitals.

JMIR Med Inform 2026;14:e108989; doi: 10.2196/108989

Keywords: new diagnosis-related group; NDRG; diagnosis coding; metadata-driven rule repository; data mart; clinical data warehouse; robotic process automation; implementation report; claim denial prevention

Introduction

Korea introduced a diagnosis-related group payment system in 1997 [1] and, in 2009, began piloting the new diagnosis-related group (NDRG) system at the National Health Insurance Service (NHIS) Ilsan Hospital—Korea's only NHIS-operated hospital and its designated test bed for health insurance policy [2]. The NDRG combines a DRG-based bundled payment with a per diem adjustment while reimbursing selected physician services separately on a fee-for-service basis. Disease group coverage expanded from 20 groups at 1 hospital to 607 groups nationwide [3,4], yet participating institutions—after peaking at 98 in 2018—declined to 90 by 2024 and 80 by January 2026 [4,5]. The decline has been attributed to unpredictable review outcomes from the Health Insurance Review and Assessment Service (HIRA) and the resulting claim denials, which also reduce hospitals' annual NDRG diagnosis coding accuracy incentives. Consequently, some institutions have concluded that continued participation no longer justifies the effort required [4].

Under DRG-based payment, diagnosis coding affects both reimbursement and the quality of national health statistics [6-8]. HIRA, Korea's public payer review agency, has conducted criteria-based reviews of NDRG diagnosis coding since 2013 [3]. HIRA's disease-specific approval criteria may differ from the coding guidelines of the Korean Classification of Diseases (KCD); for example, a clinically established diagnosis such as pleural effusion may require documented evidence of drainage or another qualifying treatment for NDRG reimbursement [3].

A more structural problem is that HIRA notifies hospitals of its review outcome only 9 to 18 months after discharge—long after the episode has closed and the claim has been paid. A subsequent denial then imposes a double penalty: (1) a direct reduction in reimbursement for that claim and (2) a reduction in the hospital's annual NDRG diagnosis coding accuracy incentive, which is calculated in part based on its denial rate. Because coders cannot know, at the time of coding, whether a diagnosis will be denied up to a year and a half later, repeated exposure to this delayed, punitive feedback teaches coders to avoid diagnoses associated with past denials, even when clinically justified—a pattern we refer to as defensive down-coding [9]. In a denial-focused review environment, this distortion may outweigh the up-coding incentives reported elsewhere [10]. Because the study hospital's own coding-finalization deadline falls only a few days after discharge—well before HIRA's feedback could be received—proactive quality management at or before the point of coding, rather than reactive correction after a denial notice, is required [11].

Clinical documentation improvement (CDI) is a well-established proactive approach, but conventional CDI relies on manual chart review by dedicated staff, limiting the coverage that can be achieved [12]. CDI programs with dedicated staff remain rare in Korea, although adoption has been growing in other health systems, including Australia [13]. Robotic process automation (RPA) has been applied to repetitive, rule-based health care tasks such as claims processing [14-16], but its application to screening for denial risk based on HIRA approval criteria is, to our knowledge, novel.

We therefore (1) operationalized HIRA's disease-specific approval criteria as data-determinable rules, (2) built a metadata-driven rule repository and a daily extract, transform, and load (ETL) pipeline that integrated 14 data sources across 6 clinical data domains into a screening data mart, and (3) implemented an RPA-based daily notification pipeline. This report describes the system's design, deployment, and the first 12 months of its operational experience.

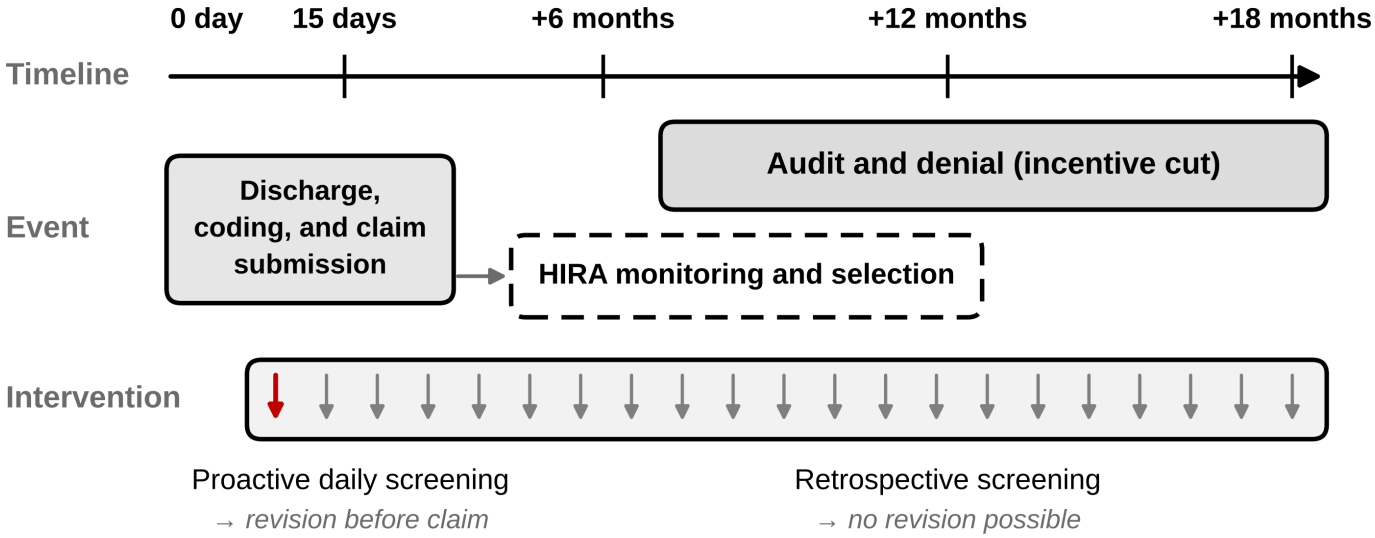
Methods

Implementation Setting and Design

The study hospital is an 820-bed general hospital that has participated as a NDRG pilot institution since 2009, with NDRG applied to approximately 80% of all inpatients. During the study period, 22,970 episodes were subject to NDRG, with approximately 93 discharge episodes requiring diagnosis coding per working day (22,970 episodes over approximately 247 working days). The Department of Medical Information Management is responsible for diagnosis coding, quality management, and clinical data analysis, with 5 health information management (HIM) professionals performing coding and quality management duties.

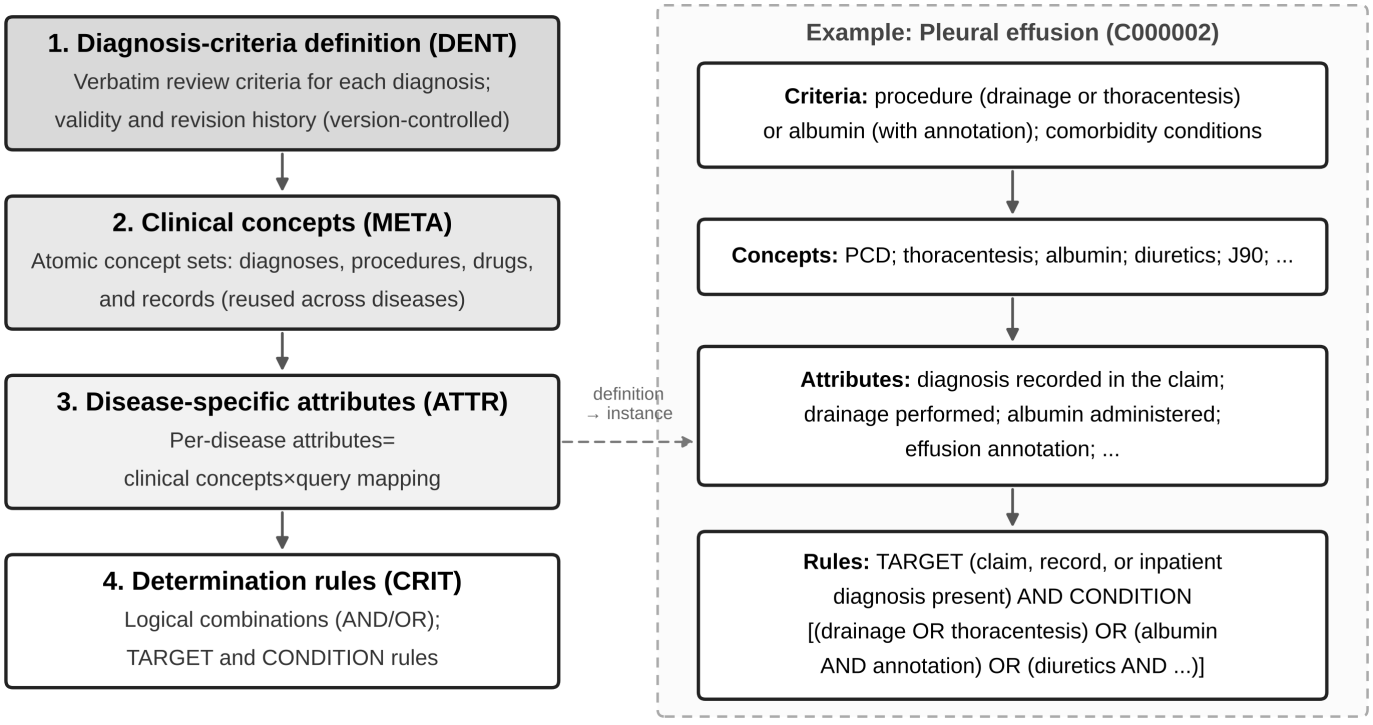
Here, "system development" encompasses the full process from criteria analysis to rule design and technical implementation, rather than software coding alone, and proceeded in 3 stages: analysis of denial history and HIRA notices to define data-determinable review criteria (stage 1); conversion of these criteria into a 4-layer rule repository and construction of the ETL pipeline and screening data mart (stage 2); and development of the RPA-based review notification pipeline (stage 3). Development ran from January 2025 to June 2025; operation and evaluation began in July 2025 and continued through June 2026. Figure 1 illustrates the key timeline for NDRG claims and the rationale for proactive, preclaim intervention rather than reactive, denial-triggered management.

Figure 1. Potential for improving coding quality according to the timing of quality management interventions along the claim review and denial timeline. HIRA: Health Insurance Review and Assessment Service.



In stage 1, the department's HIM professionals reviewed 3 years of HIRA denial notices and internal review records for each candidate diagnosis to identify the specific documentary or procedural elements HIRA required for approval. For example, denial history analysis showed that pleural effusion claims were frequently denied when the medical record lacked documentation of a definitive procedure (drainage or thoracentesis), despite a clinically established diagnosis. This finding directly informed the corresponding CONDITION rule, which requires evidence of drainage, thoracentesis, or an equivalent clinical indicator (System Architecture and Rule Repository, Figure 2). The same denial pattern analysis was applied systematically across all candidate diagnoses to determine the clinical concepts (META) and attributes (ATTR) required for each disease's rule set.

Figure 2. Four-layer structure of the review-criteria rule repository. J90 refers to the Korean Standard Classification of Diseases code for pleural effusion. PCD: percutaneous catheter drainage.



Screening targets were selected using three criteria: (1) a history of frequent denials or review issues over the preceding 3 years; (2) coder variation arising from a discrepancy between HIRA and KCD diagnosis criteria; and (3) the existence of an objective, data-determinable indicator. The resulting monitored categories and diagnoses—frequently denied diagnoses, frequent present-on-admission (POA) coding errors, and the NDRG cost-outlier group—are shown in Table 1.

Table 1. Monitored diagnoses and categories[a].

Categories and target diagnosis	Unit of analysis
Frequently denied diagnoses	
Pleural effusion	Diagnosis
Hyponatremia	Diagnosis
Neutropenia	Diagnosis
Acute lower respiratory infection (without chronic obstructive pulmonary disease)	Diagnosis
Nontraumatic vertebral fracture	Diagnosis
Moderate-or-worse malnutrition	Diagnosis
Acute kidney injury	Diagnosis
Frequent POA[b] coding errors	
Pressure injury	Diagnosis
Pseudomembranous colitis	Diagnosis
Diagnoses presumed to be POA (eg, chronic diseases)	Diagnosis
NDRG[c] cost-outlier group: suspected missing diagnosis	Admission-to-discharge episode

[a]NDRG cost-outlier group: the fee-for-service–equivalent cost exceeds the NDRG reimbursement by ≥2 million Korean won (1 KRW=US $0.00074, as of November 2, 2026); the difference is separately reimbursed by the Health Insurance Review and Assessment Service.
[b]POA: present-on-admission.
[c]NDRG: new diagnosis-related group.

Following technical development, the system was integrated into routine coding practice through a shared review-tracking workflow. Each working day at 8:30 AM, RPA compiles candidates, notifies the responsible coder by email, and appends cases to a shared Excel file (version 365; Microsoft Corp). Coders review the candidates, revise coding when warranted, and record the outcomes. A supervisor reviews the entries once daily and transfers finalized outcomes to the final result table (for more information, refer to the System Architecture and Rule Repository section). A dedicated structured review-entry interface has not yet been developed and is planned as a future improvement.

This implementation report was prepared in accordance with the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) [17] (Checklist 1).

Data Governance and Security

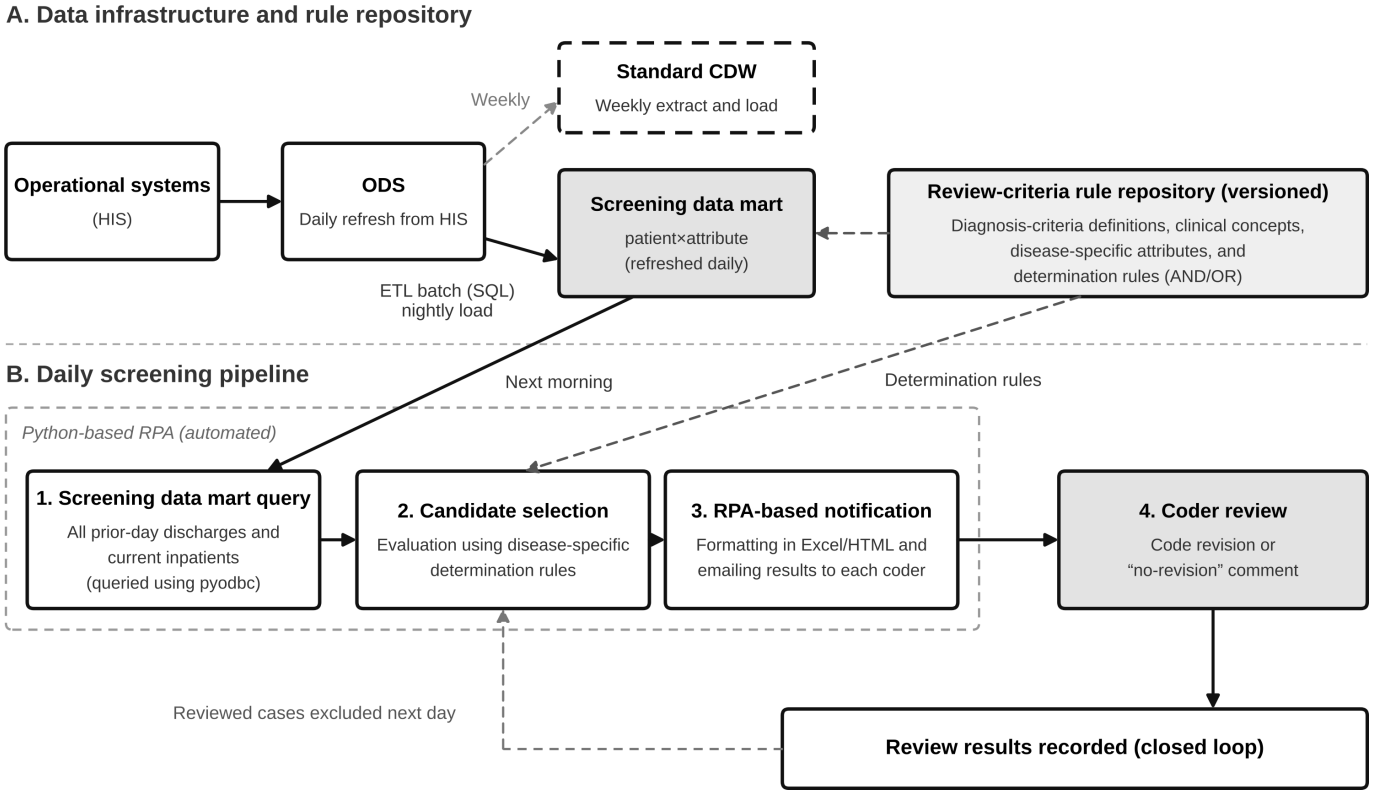
The screening system operates within the hospital's existing data governance structure. Authorized HIM professionals curate screening data and have the necessary access and write privileges within the clinical data warehouse (CDW), while database administration and infrastructure remain under the hospital's information system governance. Database access requires individual account authentication, and account-level access is logged and subject to periodic audit.

Source data are transferred through established data pipelines rather than queried directly from operational clinical databases. The rule repository, screening data mart, and results are hosted on premises within the hospital's segmented internal network, which is protected by firewalls and reachable only via a virtual private network (VPN), with access restricted to authorized personnel. The daily review-tracking file, which contains patient identifiers needed for coding review, is maintained on a designated computer within the Department of Medical Information Management and is not distributed beyond authorized department staff.

System Architecture and Rule Repository

The hospital's data infrastructure combines a daily-refreshed operational data store (ODS) with a CDW refreshed only weekly—too slow to deliver a same-morning review list. We therefore built a dedicated screening data mart within the CDW, refreshed nightly directly from the ODS (Figure 3). The mart is a single integrated structure with a disease category column, allowing disease-specific determination rules to operate on a shared mart rather than on separate marts for each disease.

Figure 3. Overview of the screening data infrastructure (Operational Data Store [ODS], clinical data warehouse [CDW], and the review-criteria rule repository) and the daily robotic process automation (RPA)–based notification pipeline. HIS: hospital information system; ETL: extract, transform, and load.



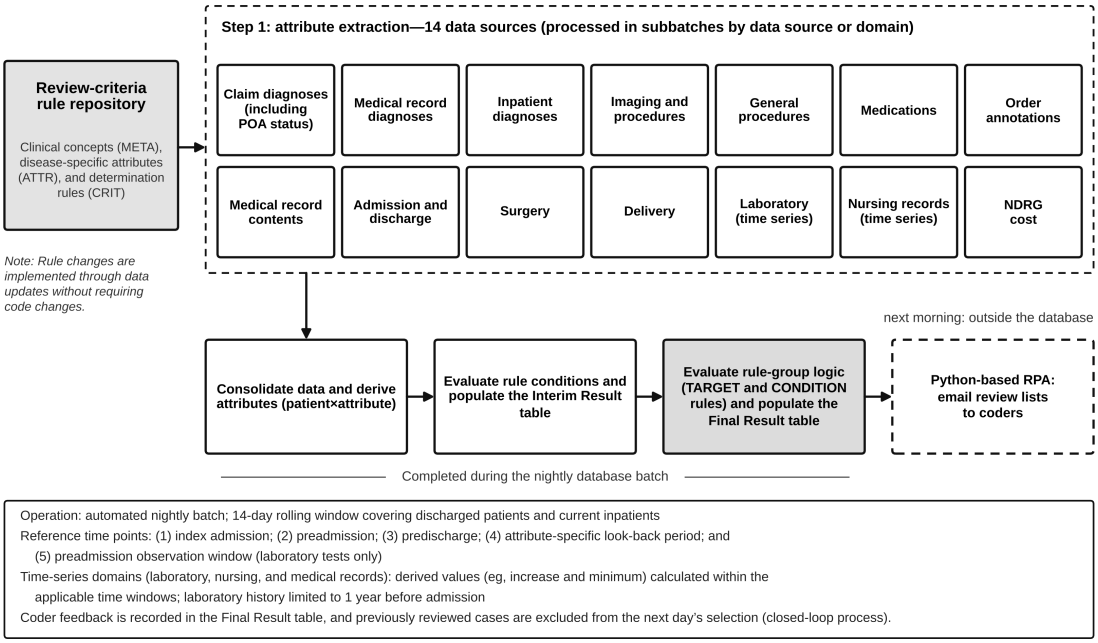
The nightly mart-loading ETL batch runs in 4 stages (Table 2; Figure 4).

Table 2. Four-stage extract, transform, and load (ETL) pipeline for the screening data mart.

Stages	Names	Key inputs	Core operations	Output
1	Target definition	ODS[a] admission and discharge records	Build (1) a rolling reference admission set comprising episodes whose admission or discharge date falls within the most recent 14 days, which is fully reloaded at each run to accommodate the ODS load cycle; and (2) a patient-level history set integrating each patient's entire past inpatient and outpatient history to determine preadmission conditions and medication history	Reference admission set; patient history set
2	Primary attribute extraction	Fourteen data sources across 6 clinical data domains (diagnoses; procedures and surgery; medication; laboratory tests; records; and admission, discharge, or cost data)	Process subbatches for each data source or domain to match candidate records using IN, BETWEEN, and LIKE operations against the matching axis defined by the clinical concepts (META); classify records according to the applicable reference time point (same-day admission, preadmission history, predischarge history, an attribute-specific look-back period, or a preadmission observation window for laboratory tests only); and apply source-specific data-cleansing rules (excluding canceled or deleted records and add-on claims, filtering by test execution and result status, selecting the latest result or record version, and restricting nursing records to those with electronic signatures)	Patient×attribute raw values
3	Time-series derivation	Time series of the same test or record item (laboratory, nursing, and other records)	Perform self-join across the time series of the same item to compute derived values, such as the magnitude of creatinine elevation used to determine acute kidney injury, and limit the historical look-back period to 1 year before admission to ensure clinical validity and query efficiency	Derived attributes
4	Integration and determination	All domain attributes generated from stages 1 to 3	Join data at the patient level; generate second-order derived attributes through arithmetic operations (difference, ratio, minimum, and maximum); evaluate the individual condition expressions in the determination rules (→ interim result table), followed by rule-group-level logic: TARGET rules (screening eligibility) and CONDITION rules (approval conditions met; → final result table)	Daily review candidates

[a]ODS: Operational Data Store.

Figure 4. Nightly extract, transform, and load (ETL) pipeline for the screening data mart (attribute extraction, rule evaluation, and robotic process automation [RPA] notification). NDRG: new diagnosis-related group; POA: present-on-admission.



Domain-specific matching used the relevant coding or identifier system for each source (eg, KCD codes for diagnoses, fee schedule codes for procedures, drug classifications or codes for medications, and test or record identifiers for laboratory and documentation data). Cost-outlier status was determined by comparing the aggregate fee-for-service–equivalent cost of all orders and prescriptions with the corresponding NDRG-calculated cost for the same episode.

Because HIRA's disease-specific approval criteria are numerous and change over time following review audits or appeals (for more information, refer to the Lessons Learned section), we stored the determination logic as data in a 4-layer rule repository (Table 3; Figure 2) rather than hard-coding it into program logic. This approach allows criteria to be updated and versioned without redeploying code.

Table 3. Four-layer structure of the review-criteria rule repository.

Layers	Table (abbreviation)	Definition	Note
1	Diagnosis-criteria definition (DENT)	Original text of the HIRA approval criteria for each managed diagnosis	Validity status and revision reason are tracked to preserve earlier versions when a criterion is revised following a review audit or appeal.
2	Clinical concept (META)	Smallest-unit clinical concepts used for determination (eg, diagnosis codes, procedure or service codes, medications, and medical record items), defined as code sets and SQL condition expressions	Each concept is registered once and reused across the rules for multiple diseases (eg, "percutaneous drainage" and "diuretic prescription").
3	Disease-specific attribute (ATTR)	Attributes required for determining each managed diagnosis, with mappings to clinical concepts and lookup queries	—[a]
4	Determination rule (CRIT)	Combination of attributes and logical operators (AND and OR) that form TARGET and CONDITION rules	TARGET rules determine screening eligibility, whereas CONDITION rules determine whether approval conditions are met.

[a]Not applicable (no additional note).

Determination results are stored in 2 linked tables: the interim result table records fulfillment status and supporting attribute values for each patient-by-rule-condition combination, while the final result table records the rule-group-level outcome, routing information, and coder review comment.

Each morning, Python (version 3.13.3; Python Software Foundation)-based RPA queries the final result table and sends candidates with supporting attribute values to the responsible coder. Finalized review outcomes are returned to the final result table, preventing reviewed cases from being selected again. RPA runs on weekdays, with accumulated weekend and holiday cases processed on the next working day.

For replication, we consider four functional components essential: (1) a structured, version-controlled repository separating clinical concepts, disease-specific attributes, and determination rules; (2) access to routinely collected clinical and administrative data; (3) a periodically refreshed data pipeline that executes the screening rules; and (4) a mechanism for delivering candidates to human reviewers and recording review outcomes. The specific technologies used here—including the ODS-CDW environment, SQL, Python,

and RPA-based email notification—are local implementation choices rather than mandatory components.

The architecture was developed specifically for this implementation rather than adapted from an existing commercial or open-source solution, and its design reflects the hospital's existing ODS-CDW environment and NDRG coding workflow.

Outcome Measures and Analysis

We report the rule repository's construction status (Table 4); coverage, defined as the number of targets evaluated relative to the hospital's total NDRG episode volume during the study period; and detection outcomes by monitored category, including candidate volume and precision, defined as the proportion of review candidates confirmed by coders as actual errors (equivalent to positive predictive value). Ninety-five percent CIs for precision were calculated using the Wilson score method because several monitored categories had relatively small numbers of review candidates or confirmed errors, resulting in proportions close to zero. The Wilson method provides more reliable interval estimates than the conventional Wald method under these conditions.

Table 4. Rule repository construction status.

Purpose and table (abbreviation)	Registered data
Diagnosis approval criteria definition	
Diagnosis-criteria definition (DENT)	11 definitions
Clinical concepts (META)	271 concepts and 5071 metadata entries
Error (up and down) likelihood determination	
Diagnosis-specific determination attributes (ATTR)	77 attributes
Determination rules (CRIT)	96 rules

Because this compared categorical outcome proportions across 3 independent, mutually exclusive categories, we used a chi-square test of independence to examine whether the probability that a review candidate was confirmed as an actual coding error differed across the 3 monitored categories (frequently denied diagnoses, POA coding items, and the NDRG cost-outlier group). The null hypothesis was that the proportion of confirmed coding errors among review candidates was equal across the 3 categories. All analyses were performed in Python (version 3.13.3), using *statsmodels* (version 0.14.6) for Wilson score CIs and *SciPy* (version 1.17.1) for the chi-square test.

Sustainability and Maintenance

Sustainability was supported by embedding the system within existing organizational and data infrastructure. The rule repository and daily screening workflow remain part of the Department of Medical Information Management's routine coding quality activities. HIM professionals review and update HIRA criteria, refine rules, and add monitoring targets as needed—in response to criterion changes or observed performance issues rather than according to a fixed schedule—while database infrastructure and technical administration are supported through existing information system governance. Versioned metadata allows most criterion changes to be implemented without redevelopment of the underlying pipeline. The repository is further supported by detailed structural documentation. Although formal handover to another developer has not yet occurred, independent data extraction and analysis for this evaluation—conducted by a department member other than the original developer—identified and corrected several metadata errors, indicating that the documentation supports use beyond the original developer.

Ethical Considerations

The screening system was developed and operated as part of routine departmental quality management activities. This report's retrospective analysis of the resulting 12-month operational data used pseudonymized, patient-level secondary data from the hospital information system and was approved by the institutional review board of NHIS Ilsan Hospital (NHIMC2026-07-002) after completion of the evaluation period. This board also waived the requirement for individual informed consent.

Implementation (Results)

Rule Repository and Coverage

As of July 2026, the rule repository contained 11 diagnosis approval criteria definitions, 271 clinical concepts (5071 associated metadata entries), 77 disease-specific determination attributes, and 96 determination rules (Table 4).

The system evaluated 17,138 diagnosis- or episode-level targets across 15,388 distinct episodes after deduplication, representing 67% of the hospital's 22,970 NDRG episodes. Coverage was calculated using the deduplicated episode count because targets were counted at the diagnosis level for some categories and at the episode level for others (Table 1).

Error Detection Outcomes

Of the 17,138 targets, 576 (3.4%) were flagged as review candidates, and 50 (8.7%, 95% CI 6.6%-11.3%) of these candidates were confirmed by coders as actual coding errors (Table 5). The proportion of review candidates confirmed as actual coding errors differed significantly across the 3 monitored categories (χ^2_2=39.9; P<.001), with precision values of 3.6% for frequently denied diagnoses, 12.7% for the NDRG cost-outlier group, and 24% for POA coding items.

Table 5. Review candidates and precision by error category[a].

Categories	Candidates evaluated, n/N (%)	Precision[b], n/N (%; 95% CI)[c]
Frequently denied diagnoses	360/2277 (15.8)	13/360 (3.6; 2.1-6.1)
New diagnosis-related group cost-outlier group	134/134 (100)	17/134 (12.7; 8.1-19.4)
Present-on-admission	82/14,727 (0.6)	20/82 (24.4; 16.4-34.7)
Total	576/17,138 (3.4)	50/576 (8.7; 6.6-11.3)

[a]The proportion of confirmed coding errors among review candidates differed significantly across the 3 categories (χ^2_2=39.9; P<.001; chi-square test of independence).
[b]Precision (%)=(revisions/candidates)×100.
[c]The 95% CIs were calculated using the Wilson score method.
[d]
[e]

Among frequently denied diagnoses, 360 of 2277 targets were flagged, of which 13 (3.6%) were confirmed as coding errors. Precision varied substantially by diagnosis, ranging from 2% (2/91) for pleural effusion and 2.4% (4/165) for acute kidney injury to 33% (1/3) for moderate-or-worse malnutrition. The cost-outlier group yielded 17 (12.7%) confirmed errors among 134 candidates, while POA screening yielded 20 (24%) confirmed errors among 82 candidates.

Implementation Resources

Development and operation of the system required no additional budget. Initial development was led by 1 HIM professional over approximately 6 months (January 2025 to June 2025) while the professional concurrently performed routine departmental duties, with coding staff participating in rule validation as part of their routine work. No personnel were hired, and no external funding, software licensing, hardware purchases, or outsourced development costs were incurred. The system runs on the hospital's existing ODS-CDW infrastructure, and growing access to AI-assisted code generation tools helped lower the technical barrier to building the SQL-Python pipeline without dedicated IT support. Because each candidate is delivered together with the supporting attribute values used in the determination (eg, laboratory results and procedure records), coders can verify most candidates without conducting a full chart review. The exception is the NDRG cost-outlier category, which requires a review of the entire medical record.

Ongoing coder review time was not prospectively recorded. Based on coders' estimates, most review candidates required approximately 1 minute of review, whereas acute kidney injury candidates, which required reviewing laboratory trends, required approximately 5 to 10 minutes, and NDRG cost-outlier candidates required approximately 10 to 20 minutes. Applying these estimates to the 576 review candidates generated over approximately 247 working days (about 2 candidates per working day) yielded an estimated approximately 40 to 80 hours of coder review time over the 12-month period, or approximately 1 to 1.5 hours per week across the department's coding staff.

Lessons Learned

Three practical lessons emerged from the first 12 months of routine operation. First, review criteria should be maintained as versioned data rather than embedded in program code. During the evaluation period, 2 cases—one involving hyponatremia and the other involving pleural effusion—were denied under stricter HIRA standards than those encoded when the cases were originally screened. Because the determination logic was stored in the rule repository, the revised criteria could be incorporated as new versioned entries without modifying or redeploying program code, while preserving the criteria applied to earlier cases. This experience demonstrated that version control is not merely a technical design feature but an operational requirement in an environment where payer review criteria can change over time.

Second, successful automation requires attention to the entire review workflow, not only to automated case detection. Although candidate identification and notification were automated, physician queries remained manual, coder review outcomes were initially recorded in a shared Excel file and subsequently transferred to the final result table, and weekend or holiday cases required an explicit catch-up rule. These operational gaps indicate that the last-mile processes connecting automated screening to human review and feedback should be designed alongside the screening algorithm itself.

Finally, screening rules require iterative refinement based on operational performance. Precision varied substantially across monitored rules, with particularly low values for acute kidney injury and pleural effusion. Rather than treating the initial rule set as fixed, we used these differences to identify rules requiring further review of their clinical concepts, attributes, or decision thresholds. Thus, implementation should be viewed as a continuous cycle of rule deployment, performance monitoring, and rule refinement rather than as a 1-time technical deployment.

Discussion

The principal achievement of this implementation is a shift from retrospective, sample-based monitoring—waiting 9 to 18 months for HIRA's review outcome—to proactive, daily screening at the point of discharge or during hospitalization, consistent with prior reports on the use of RPA to automate rule-based health care tasks [14-16,18]. Its core originality lies in converting HIRA's approval criteria—which require

evidence of a definitive treatment in addition to a clinically established diagnosis, such as a defined magnitude of creatinine elevation for acute kidney injury or drainage for pleural effusion [3]—into data-driven determination queries applied to already-structured CDW data. This approach aligns methodologically with prior work that derived clinically meaningful determination indicators from structured data alone [19] and differs from CDI approaches that rely on natural language processing (NLP) of unstructured text [12,20]. Unlike traditional CDI, the system does not itself query physicians to improve documentation; instead, it screens already-available structured data and routes candidates to coders, who may still query physicians manually as needed. Nevertheless, it shares CDI's proactive, preclaim philosophy. [11,21].

The overall precision of the screening system was modest (8.7%), indicating that most flagged cases did not ultimately require a coding revision. This finding underscores that the system should be regarded as a screening and prioritization tool rather than an autonomous error detection system; all flagged cases require coder review before any coding change is made. Nevertheless, the system narrowed 17,138 evaluated targets to 576 (3.4%) review candidates, concentrating manual review on a substantially smaller subset. The practical trade-off is that false-positive alerts still impose additional review workload and may contribute to alert fatigue. Although review time was not prospectively measured, coder-reported estimates suggest a modest burden of approximately 1 to 1.5 hours per week (for more information, refer to the Implementation Resources section). The net effect on coder workload, including the time saved through earlier correction, therefore warrants prospective evaluation.

Precision varied substantially across categories, indicating that the overall value of 8.7% should not be interpreted as representing uniform performance across rules. Category- and rule-level precision can instead guide iterative refinement by identifying low-yield rules while retaining those that generate more actionable candidates.

This report has several limitations. The architecture may be transferable to other NDRG hospitals, but local adaptation is required because database structures, available data, coding workflows, case mix, and coding practices differ across institutions. Accordingly, the precision observed here should not be assumed to generalize, and multi-institutional validation is required. HIRA's formal review for the study period had not concluded at the time of writing, so sensitivity and specificity could not be assessed; this report therefore relies on precision as the primary detection metric. The descriptive implementation design does not examine factors associated with precision through inferential statistics. Beyond coder-reported time estimates, workload and alert fatigue were not formally assessed, although no such issues were spontaneously reported during the 12-month operational period. Finally, because the department that designed and operates the system also confirmed flagged candidates as coding errors, the evaluation lacks independent or blinded verification and may be subject to evaluator bias; external validation is a priority for future work.

Planned follow-up will evaluate the system against formal HIRA review outcomes from the NDRG coding accuracy evaluation. At the time of writing, HIRA review outcomes were available for approximately the first 7 months of the postimplementation period. These outcomes are monitored weekly by the Department of Medical Information Management and are already being used to refine screening rules, with outcomes for an additional 1 to 2 months expected by November 2026. We therefore plan an initial pre-post evaluation in December 2026, comparing approximately 9 months of postimplementation discharges (July 2025 to March 2026) with discharges in calendar year 2024. The development period (January 2025 to June 2025), during which screening criteria were refined and piloted, will be excluded as a transition period. The evaluation will be conducted by the Department of Medical Information Management (DS and JH), which is also the hospital's designated department for clinical data services. HIRA review outcomes and claims data will be extracted through the hospital's formal data use application and approval procedure. Primary outcomes will be the denial rate for monitored diagnoses (denied claims per reviewed claim) and, as an indicator of defensive down-coding, the coding frequency of monitored diagnoses per 1000 NDRG episodes. Pre-post differences will be estimated as rate ratios with 95% CIs using Poisson regression adjusted for calendar month to account for seasonal variation in case mix. In addition, each postimplementation denial will be matched against the screening results to determine whether the corresponding case had been flagged, providing a preliminary estimate of rule-level sensitivity for HIRA-denied cases. These results will be updated once HIRA review of the full 12-month postimplementation period is complete, which is expected by the end of 2027. Longer-term development will explore linking structured data screening with NLP-based analysis of physician documentation and AI-assisted International Classification of Diseases (ICD) coding [12,22].

In conclusion, this metadata-driven rule repository and daily RPA-based screening system demonstrates a practical approach to transitioning from retrospective, sample-based denial management to proactive coding quality management, using existing departmental staff and infrastructure and requiring no additional budget. The system provides coders with timely, data-based evidence before claim submission; however, its performance may vary across institutions depending on local data structures, coding practices, and case mix. Further multi-institutional evaluation is needed to establish its transferability and performance in other NDRG hospitals.

Generative AI Use

The authors declare that they used Claude Sonnet 5 (Anthropic PBC) to assist with English-language proofreading and editing and Korean-English translation of drafting materials. The AI was not used to generate scientific content, figures, tables, or references. All AI-assisted text was reviewed, verified, and edited by the authors, who take full responsibility for the content.

Funding

The authors declare that no external funding was received for this work. The staff time and infrastructure used for development and operation are described in the Implementation Resources section.

Data Availability

The datasets analyzed during this study consist of pseudonymized, patient-level clinical and claims information and are not publicly available due to institutional data governance policies and patient privacy protections. Aggregated data supporting the findings of this study may be made available by the corresponding author upon reasonable request and with approval from the National Health Insurance Service Ilsan Hospital Institutional Review Board.

Authors' Contributions

DS conceived and designed the study, directed and coordinated the implementation, performed the analysis, and drafted and revised the manuscript. JH developed the rule repository and screening system.

Conflicts of Interest

DS, as head of the department operating the system, conducted a self-evaluation of the system. The other author declares no other conflicts of interest.

Checklist 1

iCHECK-DH checklist.
[DOCX File (Microsoft Word File), 16 KB-Checklist 1]

References

1. Shin HW, Yeo NG, Lee CS, Kang GW, Oh SJ, Lee JE. Evaluation of the new comprehensive payment system pilot project [Report in Korean]. Health Insurance Review and Assessment Service; 2020. URL: https://repository.hira.or.kr/handle/2019.oak/2273 [Accessed 2026-11-03]
2. New DRG payment system pilot project guideline (January 2025) [Guideline in Korean]. Health Insurance Review and Assessment Service. 2025. URL: https://repository.hira.or.kr/handle/2019.oak/3461 [Accessed 2026-11-03]
3. Shin DG, Choi MY, Choi ES, et al. New DRG diagnosis coding monitoring system improvement research [Report in Korean]. Health Insurance Review and Assessment Service; 2020. URL: https://repository.hira.or.kr/handle/2019.oak/2522 [Accessed 2026-11-03]
4. Kwak SS. "New comprehensive payment system pilot project, with many areas for improvement, should be terminated" [Article in Korean]. Doctors News (Cheongnyeon Uisa). Jun 23, 2025. URL: http://www.docdocdoc.co.kr/news/articleView.html?idxno=3029445 [Accessed 2026-09-08]
5. Ministry of Health and Welfare, Health Insurance Review and Assessment Service, National Health Insurance Service. New DRG payment system pilot project guideline [Guideline in Korean]. Health Insurance Review and Assessment Service; 2026.
6. Kim SI, Kim SJ. Understanding and evaluation of the new comprehensive DRG payment system [Article in Korean]. Health Policy Forum. 2017;15(4):44-50. URL: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07302849 [Accessed 2026-11-04]
7. Campbell SE, Campbell MK, Grimshaw JM, Walker AE. A systematic review of discharge coding accuracy. J Public Health Med. Sep 2001;23(3):205-211. [doi: 10.1093/pubmed/23.3.205] [Medline: 11585193]
8. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med. Feb 11, 1988;318(6):352-355. [doi: 10.1056/NEJM198802113180604] [Medline: 3123929]
9. Pongpirul K, Robinson C. Hospital manipulations in the DRG system: a systematic scoping review. Asian Biomed. 2013;7(3):301-310. [doi: 10.5372/1905-7415.0703.180]
10. Zhou W, Xu C, Zhang L, Fu H, Jian W. Behaviours and drivers of diagnosis-related group upcoding in China: a mixed-methods study. Soc Sci Med. Feb 2025;366:117660. [doi: 10.1016/j.socscimed.2024.117660] [Medline: 39721170]
11. Adepoju KO. Review of the impact of clinical documentation improvement (CDI) programs on coding accuracy, reimbursement, and quality of care. Unizik J Educ Res Policy Stud. 2024;18(1). URL: https://unijerps.org/index.php/unijerps/article/view/719 [Accessed 2026-11-01]
12. Perkins SW, Muste JC, Alam T, Singh RP. Improving clinical documentation with artificial intelligence: a systematic review. Perspect Health Inf Manag. 2024;21(2):1d. [Medline: 40134899]

13. Hay P, Wilton K, Barker J, Mortley J, Cumerlato M. The importance of clinical documentation improvement for Australian hospitals. Health Inf Manag. Jan 2020;49(1):69-73. [doi: 10.1177/1833358319854185] [Medline: 31303061]
14. Nimkar P, Kanyal D, Sabale SR. Increasing trends of artificial intelligence with robotic process automation in health care: a narrative review. Cureus. Sep 2024;16(9):e69680. [doi: 10.7759/cureus.69680] [Medline: 39429258]
15. Chaudhary RL, Peddireddy AR, Mathur U. The impact of robotic process automation (RPA) in high-dollar claims processing in healthcare. Eur J Adv Eng Technol. 2024;11(6):47-49. [doi: 10.13140/RG.2.2.28723.23847]
16. Park A, Jung SY, Yune I, Lee HY. Applying robotic process automation to monitor business processes in hospital information systems: mixed method approach. JMIR Med Inform. Mar 7, 2025;13:e59801. [doi: 10.2196/59801] [Medline: 40053771]
17. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res. May 10, 2023;25:e46694. [doi: 10.2196/46694] [Medline: 37163336]
18. Huang WL, Liao SL, Huang HL, et al. A case study of lean digital transformation through robotic process automation in healthcare. Sci Rep. Jun 25, 2024;14(1):14626. [doi: 10.1038/s41598-024-65715-9] [Medline: 38918486]
19. Xiang J, Durance PW, Griffes LC, Chen Y, Bakshi RR. Measuring case severity: a novel tool for benchmarking and clinical documentation improvement. BMC Health Serv Res. Apr 15, 2022;22(1):513. [doi: 10.1186/s12913-022-07935-1] [Medline: 35428299]
20. Rouse M, Jones M, Zogleman B, May R, Ekilah T, Gibson C. Resident integration with inpatient clinical documentation improvement: a quality improvement project. BMJ Open Qual. Jun 2022;11(2):e001300. [doi: 10.1136/bmjoq-2020-001300] [Medline: 35732406]
21. Rodenberg H, Shay L, Sheffield K, Dange Y. The expanding role of clinical documentation improvement programs in research and analytics. Perspect Health Inf Manag. 2019;16(Winter):1d. [Medline: 30766455]
22. Kaur R, Ginige JA, Obst O. AI-based ICD coding and classification approaches using discharge summaries: a systematic literature review. Expert Syst Appl. Mar 2023;213:118997. [doi: 10.1016/j.eswa.2022.118997]

Abbreviations

CDI: clinical documentation improvement
CDW: clinical data warehouse
ETL: extract, transform, and load
HIM: health information management
HIRA: Health Insurance Review and Assessment Service
ICD: International Classification of Diseases
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
KCD: Korean Classification of Diseases
NDRG: new diagnosis-related group
NHIS: National Health Insurance Service
NLP: natural language processing
ODS: operational data store
POA: present-on-admission
RPA: robotic process automation
VPN: virtual private network

Edited by Caroline Perrin; peer-reviewed by Awa Babington-Ashaye; submitted 07.Aug.2026; initial comments to author 04.Sep.2026; final revised version received 01.Oct.2026; accepted 20.Oct.2026; published 20.Nov.2026

Please cite as:
Shin D, Hwang J
Development and Evaluation of a Metadata-Driven Rule Repository and Daily Screening System for Diagnosis Coding Compliance Under the Korean New Diagnosis-Related Group Payment System: Implementation Report
JMIR Med Inform 2026;14:e108989
URL: https://medinform.jmir.org/2026/1/e108989
doi: 10.2196/108989

© Donggyo Shin, Jinhee Hwang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.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.