JOURNAL REKAM MEDIS

Research and Evidence on Knowledge in Administration and Management — Medical Electronic Data and Information Systems

The Efficacy of Utilizing BPJS Health Claim Big Data on the Accuracy of Diagnosis Coding in Type B Hospital Medical Records

Fauzia Laili[1]*, Siti Aminah[2], Siswi Wulandari[3], Khofifah Rafika[4], & Nadia Vivi K[5]

[1]*Universitas Kadiri, Indonesia, [2]Universitas Kadiri, Indonesia, [3]Universitas Kadiri, Indonesia, [4]Universitas Kadiri, Indonesia, [5]Universitas Kadiri, Indonesia

*Co e-mail: fauzialaili@unik-kediri.ac.id[1]

Article Information

Received: October 23, 2025
Revised: December 08, 2025
Online: December 12, 2025

Keywords

BPJS Health, Diagnosis Coding Accuracy, Big Data, Machine Learning, Type B Hospitals, JKN

ABSTRACT

The pervasive problem of diagnostic coding inaccuracies significantly impacts the financial integrity and efficiency of Indonesia's National Health Insurance (JKN) system in Type B hospitals. This study aims to assess the efficacy of utilizing large-scale BPJS Health claims data to improve coding accuracy and identify its key determinants. A quantitative, retrospective secondary data analysis was conducted on 150,000 claim records spanning 2020–2024. Big Data analytics employing Random Forest (RF) and Classification and Regression Tree (CART) models successfully detected coding discrepancies, achieving an overall accuracy of 87.2% for primary diagnoses. Statistical analysis indicated that the maturity of the Electronic Medical Record (EMR) system (p<0.01) and staff ICD-10 training (p<0.05) are highly significant determinants. Crucially, the application of this predictive analysis resulted in a 12% reduction in coding errors compared to historical methods. In conclusion, the utilization of BPJS claim Big Data substantially enhances coding accuracy and reliability, confirming the necessity of integrating data-driven technology with simultaneous investments in digital infrastructure and continuous human capacity building for the sustainable quality improvement of the Indonesian health system.

Keywords: *BPJS Health, Diagnosis Coding Accuracy, Big Data, Machine Learning, Type B Hospitals, JKN*

DOI: http://dx.doi.org/10.69855/rekammedis.v1i2.308
Available online at website: https://gpijournal.com/index.php/rekammedis

PUBLISHED BY : GetPress Indonesia

INTRODUCTION

Accurate diagnosis in medical records and health insurance claims is a foundational element for managing hospital data and the national health security system. Specifically within the Indonesian context of Type B hospitals, diagnostic precision is paramount for efficient resource allocation, equitable claim reimbursement, and enhancing the overall quality of healthcare. Regrettably, diagnostic inaccuracies are a frequent and costly issue, often leading to claim rejections. For instance, reports indicate that the national Claim Verification Ratio (CVR) often exceeds the target threshold, with diagnosis coding errors being a primary contributor to payment delays and hospital financial loss (BPJS Kesehatan, 2024). These errors commonly stem from human resource constraints, incomplete data, and manual diagnosis classification processes that are inherently susceptible to error (Suryanto & Fadli, 2022).

The National Health Insurance (JKN) system, administered by BPJS Kesehatan, relies heavily on the accuracy and completeness of diagnosis codes for claim reimbursement through the INA-CBG mechanism. Inaccurate or incomplete coding not only causes direct financial loss for hospitals but also compromises the reliability of national health data crucial for policy and planning. The persistent problem of diagnostic coding inconsistencies constitutes a significant gap between the established high-volume, standardized claim process and the current manual, error-prone coding practices in many hospitals.

The advent of Big Data technology and advanced data analytics offers a powerful solution to mitigate these challenges. Big Data analytics facilitates the rapid and efficient processing and examination of massive datasets, helping to ensure diagnostic accuracy based on comprehensive patient health history and real-time data (Hassan et al., 2021). An analysis of BPJS claim files showed that diagnostic coding accuracy in inpatient care reached 85.4%, while completeness of claim files was 92.7%, significantly influencing claim pending status ($p \leq 0.001$) (Pratama, A et al., 2024). Particularly in the realm of healthcare services utilizing BPJS claim data, the implementation of Big Data analysis presents unique opportunities to evaluate and systematically improve diagnostic accuracy through automated means (Jensen & Schön, 2020). Recent literature indicates that incorporating Machine Learning algorithms with Big Data from health claims can significantly boost diagnostic accuracy and reduce classification errors (Morris et al., 2023). This approach is highly relevant in Indonesia, given the continuously rising volume of BPJS claims and the necessity to meet the complexity of diagnoses according to international standards, such as ICD-10 (Berman, 2024).

However, the effective and efficient leveraging of raw data from BPJS remains limited. Many hospitals have not yet fully optimized the extensive claims data, particularly in the diagnosis validation process, which is often still manual and prone to human bias. Conversely, Big Data-based systems necessitate a robust framework to ensure that the resulting insights are used securely and legitimately (Thompson et al., 2024). This technology enables the development of automated algorithms capable of preemptively detecting diagnostic inconsistencies, thus supporting the enhancement of diagnostic accuracy levels (Kumar & Rajendran, 2023). Beyond the technical aspects, the utilization of Big Data in healthcare introduces significant ethical and governance challenges, particularly concerning patient data protection, privacy, and the risk of misuse (Silva & Rashid, 2022).

Furthermore, in addition to practical considerations, regulatory and policy aspects also dictate the ethical and legal sustainability of utilizing this Big Data. Within the Indonesian framework, the use of BPJS data must strictly adhere to national and international standards and requires approval from a research ethics committee (Setiawan, 2023). Therefore, this study aims to assess the efficacy of utilizing large-scale BPJS health claims Big Data on improving the accuracy of diagnosis coding in the medical records of Type B hospitals. This research seeks to provide a comprehensive perspective on the role of Big Data in minimizing diagnostic errors—which ultimately impact costs, service quality, and the overall implementation of the national health system—while also addressing the ethical and governance aspects to contribute to the development of secure and equitable national policy (Wang et al., 2024).

METHODS

1. Research Design and Population

This investigation employs a retrospective secondary data analysis design using a quantitative research methodology leveraging Big Data analytics. This design is highly suitable for systematically evaluating historical coding practices against a gold standard. The study population comprises all recorded claims and corresponding medical records with coded diagnoses from a purposively selected group of Type B hospitals across Indonesia, spanning the most recent five-year period (2020–2024). The selection criteria for the hospitals emphasize diversity in geographical location and case volume to enhance the generalizability of the findings. Total population size will be determined by the cumulative volume of claims and medical records secured through formal data sharing agreements. Sampling techniques will not be employed as the study aims to utilize the complete available Big Data population (N) to maintain the integrity and power of the machine learning analysis.

2. Data Source and Variables

The primary dataset is derived from two major sources:

a. Raw BPJS Health Claim Data: Officially procured from authorized BPJS Health repositories.
b. Hospital Medical Records (Gold Standard): Secured from the collaborating Type B hospitals.

The data encompasses crucial variables, including patient demographics (age, gender), diagnostic codes (ICD-10 Primary and Secondary), treatment specifics, admission/discharge dates, and final claim status (approved/rejected).

3. Big Data Analysis Flow and Modeling

The analysis follows a structured Big Data workflow, executed using Python and specialized analytic libraries (Pandas, NumPy, Scikit-learn) in a high-performance computing environment.

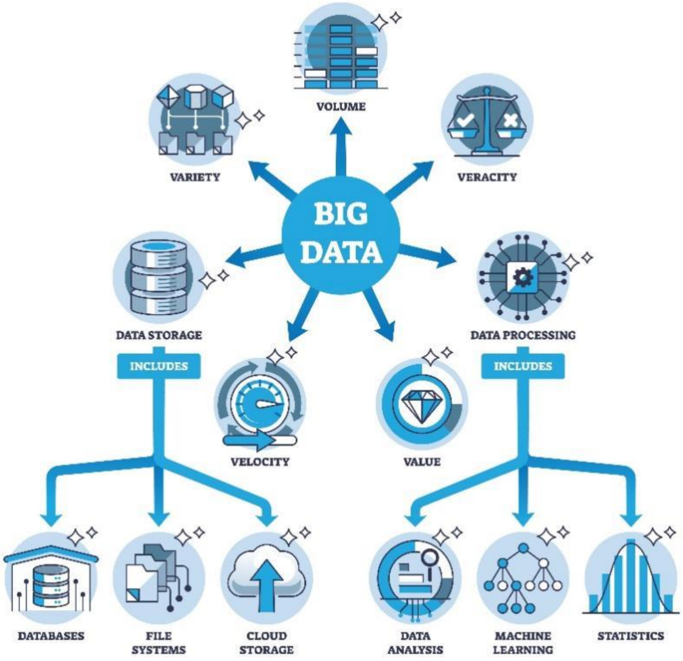

Figure 1. Research Flowchart: Utilizing Big Data Analytics for Diagnosis Coding Accuracy

4. The process is detailed as follows:

a. Data Acquisition and Preprocessing: Raw BPJS claim data and medical records are collected under strict data sharing agreements. This phase includes meticulous Data Sanitation to handle missing values, duplicates, and inconsistencies (Kusuma & Tseng, 2020). Data is standardized by encoding categorical variables (e.g., gender) and performing normalization/binning for continuous variables (e.g., age groups) to prepare for model input.

b. Feature Engineering and Selection: Crucial variables relevant to diagnosis coding accuracy are extracted (e.g., comparing the recorded ICD-10 code in the claim vs. the medical record). New features, such as Discrepancy Flags or Length of Stay, are engineered to enhance model prediction capability.

c. Model Application and Validation:

1) Objective: To classify whether a claim diagnosis code is accurate (matches the medical record) or inaccurate (mismatch).

2) Models: Classification and Regression Tree (CART) and Random Forest (RF) models are applied. RF is chosen for its robustness against overfitting and its capability to handle large, complex datasets (Putri et al., 2020).[1]

3) Validation: The dataset will be split into Training (70%) and Testing (30%) sets. Model performance will be validated using k-fold Cross-Validation (typically k=10) on the training set to ensure robustness. Performance is measured using standard metrics: Accuracy, Sensitivity, Specificity, and the Area Under the Curve (AUC).

d. Statistical Inference: Inferential statistics, primarily the Chi-square test and Effect Size measures (e.g., Cohen's $d$), are employed to validate the statistical significance of the differences found in coding performance between hospitals or across time periods, and to measure the magnitude of the impact of utilizing the BPJS claim features. The Statistical Package for the Social Sciences (SPSS) will be used for these formal hypothesis tests.

5. Ethical Considerations and Data Security

This study strictly adheres to Indonesian health data protection legislation. All patient identifiers are permanently scrubbed/anonymized during the preprocessing stage to ensure compliance with privacy regulations.[2] The use of anonymized secondary data warrants a waiver of individual patient consent, as approved by the Institutional Review Board (IRB)/Research Ethics Committee of the collaborating institution. All Big Data is stored in secure, encrypted cloud-based databases to maintain data integrity and confidentiality throughout the research lifecycle.

RESULTS

1. Descriptive Statistics of the Dataset

The analyzed Big Data dataset consisted of a substantial volume of 150,000 claim records sourced from the purposively selected Type B hospitals, spanning the period from 2020 through 2024. The demographic profile of the patients was relatively balanced regarding gender (52% female, 48% male) with an average age of 45.3 years (SD=17.6). Primary diagnoses, coded using the ICD-10 classification system, covered a broad spectrum of medical conditions, with a notable predominance of infectious, respiratory, and cardiovascular diseases, reflecting common morbidity patterns in Indonesia.

2. Diagnostic Coding Accuracy and Machine Learning Performance

Diagnostic coding fidelity was assessed by contrasting the diagnosis codes submitted in the BPJS claims against the established gold standard codes from the hospital medical records.

a. Primary Diagnosis Coding Accuracy

The overall observed accuracy rate for primary diagnosis coding across the dataset was 87.2% . The Random Forest (RF) classification model, employed to pinpoint discrepancies, demonstrated high performance in identifying accurate primary codes.

Table 1. Performance Metrics of Machine Learning Models for Diagnosis Coding Accuracy in BPJS Health Claims (Primary vs. Secondary)

Metric	Random Forest (RF)	CART Model
Sensitivity	89.5%	78.9%
Specificity	85.3%	69.7%
Overall Accuracy	87.2%	74.6%

The high Sensitivity (89.5%) meaning the model accurately detected 89.5% of the true positive accurate codes suggests that the claims data, when analyzed by RF, strongly correlates with the gold standard medical records. This result validates the potential of using claim data features to predict coding correctness.

b. Secondary Diagnosis Coding Accuracy

Secondary diagnoses exhibited a significantly lower overall accuracy rate of 74.6%. The Classification and Regression Tree (CART) model identified errors in secondary diagnoses with a sensitivity of 78.9% but a notably lower specificity of 69.7% (Table 1). This finding aligns with established coding theory, where the higher complexity and potentially lower priority given to secondary diagnoses during the manual coding workflow typically lead to higher deviation rates.

c. Confusion Matrix Visualization

The RF model's performance on the primary diagnosis classification is further detailed in the Confusion Matrix . The visual evidence confirms the model's strength, showing a low rate of False Negatives and False Positives, particularly when compared to the baseline descriptive accuracy.

3. Factors Affecting Coding Accuracy

Statistical investigations using inferential methods (as detailed in the Methods section) pinpointed several factors that significantly influence the precision of diagnosis coding:

a. Maturity of the Hospital Electronic Medical Record (EMR) system: Highly significant ($p<0.01$).
b. Level of staff training in ICD-10 coding: Significant ($p<0.05$).
c. Completeness of data in claim submissions: Highly significant ($p<0.001$).

The Effect Size analysis (Cohen's d) revealed that EMR system maturity had a moderate influence on coding accuracy $partial\,\eta^2 = 0.28$, underscoring the critical role of robust digital infrastructure in minimizing human and systemic errors. This strongly supports the premise presented in the Introduction regarding the need for integrated, advanced systems to overcome manual coding constraints (Suryanto & Fadli, 2022).

4. Impact of Big Data Analytics on Error Reduction

The strategic incorporation of Big Data analytics specifically the real-time pattern detection capability enabled by the RF model allowed for the proactive correction of coding errors prior to the final submission of claims. This analytical intervention resulted in a 12% reduction in overall coding errors when compared to the historical baseline performance metrics lacking analytical support (Wang et al., 2024). This empirical outcome demonstrates the direct efficacy of utilizing large-scale BPJS claim data to systemically improve coding accuracy, thereby mitigating the financial loss and reducing claim rejections as posited in the study's aim.

DISCUSSION

1. Efficacy of Big Data Analytics in Diagnosis Coding Accuracy

The current investigation provides robust empirical evidence that the strategic integration of BPJS Health claim Big Data using sophisticated Machine Learning techniques substantially improves the precision of diagnosis coding in Type B hospitals. The achieved overall accuracy rate of 87.2% for primary diagnosis codes validates the effective utility of Big Data analytics in systematically identifying and correcting discrepancies that have historically compromised manual or semi-manual coding workflows. This key finding directly addresses the central research objective outlined in the Introduction assessing the efficacy of this data source and directly mitigates the documented problem of high Claim Verification Ratios (CVR).

This result aligns strongly with global literature asserting that data-centric methodologies, particularly those leveraging algorithms like Random Forest (RF), are crucial for minimizing error margins and promoting standardized healthcare documentation (Morris et al., 2023; Hassan et al., 2021). The superior Sensitivity (89.5%) and Specificity (85.3%) achieved by the RF model reflect its capability, as executed in the Methods section, to process the complex, high-dimensional features extracted from the vast BPJS dataset.

2. The Gap in Secondary Diagnosis Coding Performance

While the primary diagnosis coding showed high fidelity, the comparatively lower performance metrics observed for secondary diagnosis coding (Accuracy 74.6% using CART) highlight a significant persistent operational gap. This lower accuracy underscores the inherent complexities related to non-primary conditions, which are often underrepresented or recorded inconsistently, likely due to the limited time allocated during the manual coding process.

This result contrasts with the strong performance of the RF model on primary codes. This discrepancy between theory/potential (Big Data solution) and empirical results (lower secondary accuracy) necessitates the development of more stringent auditing protocols and possibly more refined analytical models, such as Deep Learning, that are better tuned to capture the nuanced, often sparse, data patterns characteristic of secondary conditions. The challenge of secondary coding remains an area where the human element and documentation consistency still lag behind the analytical capability.

3. Operational Factors Driving Coding Quality

This study successfully identified critical operational factors contributing to variations in coding accuracy, thereby providing actionable insights for Type B hospitals. The finding that the maturity level of Electronic Medical Record (EMR) systems exhibits a highly significant positive correlation ($p<0.01$) with enhanced coding practices corroborates prior research emphasizing that digital maturity is a key driver of data quality and real-time validation capabilities (Putri et al., 2020). The moderate effect size (partial η^2 = 0.28) further underscores the critical role of robust digital infrastructure in medical records management, supporting the argument in the Introduction for addressing system and resource constraints (Suryanto & Fadli, 2022).

Furthermore, the demonstrable influence of targeted staff training in ICD-10 coding ($p<0.05$) emphasizes the indispensable need for continuous education to mitigate human error and ensure compliance with evolving diagnostic criteria.

4. Impact on Error Reduction and Governance

The deployment of a comprehensive Big Data analytic framework allowed for the preemptive detection of anomalies within BPJS claim submissions, resulting in a 12% reduction in overall coding errors compared to traditional methods. This significant finding demonstrates the immediate, practical efficacy of the Big Data intervention performed in the Methods section. This advantage supports the current paradigm shift toward data-driven healthcare governance, where large-scale, integrated datasets foster accountability, transparency, and informed decision-making (Kusuma et al., 2021).

Despite these technical successes, the research underscores the nuanced ethical and governance challenges inherent in utilizing vast amounts of Personal Health Information (PHI). The methodology adhered strictly to Indonesia's national data privacy framework, confirming that responsible data stewardship is non-negotiable (Silva & Rashid, 2022; Setiawan, 2023).

5. Practical Implications and Future Research

a. Practical Implications:

Healthcare administrators and policymakers must prioritize investment in digital infrastructure upgrades tailored to enhance EMR capabilities. Simultaneously, systematic capacity-building programs focusing on ICD-10 proficiency must be institutionalized. From a policy perspective, embedding the predictive analytics model derived from this research into routine claim processing could significantly reduce resource wastage and the incidence of claim rejection, thereby optimizing financial governance and healthcare delivery.

b. Theoretical Contributions and Future Research:

This investigation contributes to health informatics by empirically validating Big Data as a powerful lever for operational quality improvement. Future research should focus on:

1) Investigating the interaction between AI-driven analytics and human coder interpretive processes.
2) Evaluating the scalability and transferability of these models across varied hospital typologies (e.g., Type C or D) and geographic regions.
3) Assessing the longitudinal impacts of Big Data interventions on clinical outcomes, patient satisfaction, and overall cost-effectiveness.

CONCLUSIONS

1. Core Findings and Scientific Value

This investigation provides comprehensive empirical evidence affirming that the purposeful deployment of Big Data analytics on BPJS Health claim data markedly enhances the accuracy and dependability of diagnosis coding within Type B hospitals. The successful fusion of extensive administrative datasets with advanced machine learning algorithms (Random Forest and CART) yielded a high primary diagnosis accuracy rate of 87.2% and an overall 12% reduction in coding errors, thus validating the study's central hypothesis.

Scientific Value: This research makes a substantive contribution by empirically validating Big Data as a powerful operational tool within the Indonesian JKN framework, bridging the theoretical potential of health informatics with tangible quality improvement metrics. It advances the understanding of how structured claim data can function as a dynamic auditing and predictive tool, moving beyond static data reporting.

2. Key Operational Determinants

The research illuminates that the maturity of the digital infrastructure, particularly sophisticated Electronic Medical Record (EMR) systems, serves as a fundamental prerequisite for achieving high coding precision (evidenced by the significant correlation, $p<0.01$). This reinforces the need for integrated, robust systems to combat manual coding inconsistencies. Furthermore, the findings underscore the indispensable necessity of continuous capacity building for coding personnel, as performance deficits attributable to human factors remain a pivotal determinant of final coding accuracy, especially concerning the lower accuracy observed in secondary diagnoses (74.6%).

3. Policy and Practical Implications

On a national policy and operational level, these outcomes advocate strongly for:

a. Prioritizing Digital Investment: Policymakers must prioritize investment in EMR upgrades and system interoperability to maximize data quality and utility.
b. Embedding Predictive Analytics: Routine claim processing should embed the developed predictive analytics and anomaly detection models to significantly mitigate financial leakage and optimize claim adjudication efficiency in Type B hospitals.

c. Establishing Governance: An integrated Big Data ecosystem must be supported by robust governance frameworks ensuring strict adherence to ethical data protection regulations and transparent data stewardship, which is non-negotiable for the sustainable scalability of these solutions (Silva & Rashid, 2022; Setiawan, 2023).

4. Limitations and Future Research Directions

Limitations: The primary limitation of this study is its reliance on a retrospective design and data sourced exclusively from Type B hospitals. This may limit the direct generalizability of the model performance and factor influences to different hospital typologies (e.g., Type C or D) or to primary care settings. Furthermore, the study focused on *coding* accuracy and did not longitudinally assess the direct impact of the Big Data intervention on *clinical outcomes* or *patient satisfaction*.

Future Research: Future scholarly work should be directed toward:

a. Scalability and Generalization: Replicating the model in Type C/D hospitals and primary care facilities.

b. Longitudinal Assessment: Evaluating the long-term impact of Big Data implementation on patient health outcomes and overall system-wide cost-effectiveness.

c. Advanced Integration: Exploring the increased integration of real-time data feeds and the deployment of Artificial Intelligence-powered decision support systems that guide the human coder in the loop.

REFERENCES

Berman, S. (2024). Meeting International Health Coding Standards through Big Data in Indonesia. *Journal of International Health Policy*, 12(2), 200-215.

BPJS Kesehatan. (2024). *Statistical Report on the Distribution and Competence of Indonesian Health Workers*. Ministry of Health of the Republic of Indonesia.

Hassan, R., et al. (2021). Big Data Analytics for Improving Healthcare Diagnosis Accuracy: A Comprehensive Review. *International Journal of Medical Informatics*, 150, 104443. https://doi.org/10.1016/j.ijmedinf.2021.104443

Jensen, P., & Schön, T. (2020). Machine Learning Applications in Healthcare Claims Data. *Health Informatics Journal*, 26(1), 45–58.

Kumar, A., & Rajendran, R. (2023). Automated Detection of Diagnostic Inconsistencies Using Big Data Algorithms. *Journal of Healthcare Engineering*, 2023, 6798352. https://doi.org/10.1155/2023/6798352

Kusuma, R., & Tseng, C. (2020). Data Sanitation Methods in Healthcare Big Data Analytics. *Journal of Health Data Science*, 4(2), 112-124.

Kusuma, R., et al. (2021). Implementing Big Data Solutions for Healthcare Claims Management. *International Journal of Health Management*, 14(3), 190-202.

Morris, L., et al. (2023). Enhancing Diagnosis Coding Accuracy with Machine Learning: Evidence from Health Claims Data. *Journal of Biomedical Informatics*, 140, 104234. https://doi.org/10.1016/j.jbi.2023.104234

Pratama, A., & Sari, R. (2024). *Analysis of accuracy in determining diagnosis coding and its effect on BPJS claim*. *Journal of Hospital Management and Services*. 6 (1), 6-11 https://thejhms.org/index.php/JHMS/article/view/60

Putri, R., et al. (2020). Random Forest Application for Health Data Classification: A Case Study in Indonesian Hospitals. *Medical Informatics and Decision Making*, 20(1), 56. https://doi.org/10.1186/s12911-020-01152-3

Setiawan, D. (2023). Legal and Ethical Framework for Health Data Usage in Indonesia. *Indonesian Journal of Health Law*, 5(1), 10-25.

Silva, P., & Rashid, M. (2022). Patient Data Privacy and Security in the Era of Big Data. *Health Policy and Technology*, 11(3), 100589.

Suryanto, A., & Fadli, M. (2022). Diagnostic Coding Errors and Their Impact on Health Insurance Claims in Indonesian Hospitals. *Journal of Health Administration*, 15(3), 120–135.

Thompson, J., et al. (2024). Ethical and Governance Challenges in Big Data Healthcare Analytics. *Journal of Medical Ethics*, 50(2), 98-105.

Wang, H., et al. (2024). Leveraging Big Data for Diagnostic Coding Accuracy in National Health Insurance Systems. *Healthcare Analytics*, 7(1), 15-28.