DOI Link: https://doi.org/10.61586/JhURi
Vol.40, Issue.12, Part.1, Dec 2025, PP.2-19

Nationwide Predictive Modeling and Socioeconomic Disparity Analysis of Intracerebral Hemorrhage in Korea

Sung-Jun Kim[1], Bong-Hyun Kim[2]*

[1]Department of Big data Contents Convergence, Namseoul University, 91 Daehak-ro, Seonghwan-eup, Seobuk-gu, Cheonan-si, Chungcheongnam-do 31020, Republic of Korea
[2]*Department of Computer Engineering, Seowon University, 377-3, Musimseo-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, Republic of Korea
bhkim@seowon.ac.kr

Received Feb 2025 Accepted Nov 2025 Published Dec 2025

Abstract
Intracerebral hemorrhage (ICH) is a highly fatal subtype of stroke associated with substantial long-term disability and socioeconomic burden. In Korea, marked regional and socio-economic disparities in incidence and outcomes persist, yet comprehensive quantitative analyses remain limited. This study analyzed nationwide hospitalization data from the Korean National Health Insurance Service covering 2006–2022 to characterize demographic and socio-economic patterns, identify high-risk regions, and predict future incidence. Data preprocessing involved removing duplicates, standardizing categorical variables, and aggregating data by region. Descriptive statistics and multiple visualization techniques—correlation heatmaps, pair plots, boxplots and violin plots, regression plots, pie charts, and histograms—were applied to assess temporal, demographic, and spatial trends. A regression-based predictive model incorporating demographic and socio-economic factors estimated 5-year regional incidence forecasts. Results indicated higher ICH rates among adults aged ≥65 years and among low-income groups, with strong correlations between inpatient episodes and 28-day mortality (r=0.82) and between patient-years and inpatient cases (r=0.85). Forecasts indicated persistent high-risk clusters in both metropolitan and rural regions. These findings highlight the value of big data analytics in guiding targeted prevention, optimizing resource allocation, and reducing health inequities in ICH management.
Keywords: Intracerebral hemorrhage, Epidemiology, Predictive modeling, Health equity, big data analysis.

1. Introduction

Intracerebral hemorrhage (ICH) is one of the most devastating subtypes of stroke, characterized by sudden bleeding within the brain parenchyma and associated with extremely high case-fatality rates and severe long-term disability. Although it accounts for only 10–15% of all stroke cases globally, ICH disproportionately contributes to stroke-related mortality and long-term dependency. In Korea, where rapid population aging is underway and cardiovascular risk factors remain prevalent, ICH continues to pose a significant and escalating public health challenge. This burden is further exacerbated by pronounced disparities in incidence, outcomes, and access to care across regions and socio-economic groups[1].

From a public health perspective, the Korean context reveals three interrelated structural problems. First, regional disparities persist, as the distribution of stroke-ready hospitals, emergency transport systems, and diagnostic imaging resources remains uneven between metropolitan areas and rural provinces. Second, socio-economic inequalities exert a profound influence, with low-income populations often facing barriers to preventive care, delays in diagnosis, and limited access to high-quality acute treatment and rehabilitation. Third, demographic shifts—specifically, a rapidly expanding elderly population—are increasing the absolute number of individuals at elevated risk, thereby amplifying overall health system demand.

Epidemiological evidence indicates that hypertension, diabetes mellitus, smoking, and excessive alcohol consumption remain highly prevalent, especially in socio-economically disadvantaged groups. These modifiable biomedical risk factors interact with non-medical determinants such as education level, household income, and geographic accessibility of healthcare, collectively influencing both disease incidence and survival outcomes. Addressing this multifactorial problem requires an integrative approach that combines rigorous epidemiological analysis with proactive health policy design, underpinned by a strong focus on equity. The present study set out to conduct a comprehensive, data-driven investigation of ICH in Korea using 17 years of nationwide hospitalization records. The objectives were threefold: (1) to quantify the demographic and socio-economic patterns in ICH burden; (2) to identify spatial heterogeneity in incidence and mortality through geographic visualization; and (3) to develop predictive models for future regional incidence, thereby enabling targeted prevention and optimized resource allocation[2][3].

Data were obtained from the Korean National Health Insurance Service via the Public Data Portal, encompassing all hospital admissions for ICH between 2006 and 2022. Each record contained detailed demographic variables (year, sex, five-year age group, income quintile), administrative region identifiers (province/metropolitan city, district/county), and clinical utilization measures including patient-years, actual patient counts, inpatient episodes, and deaths within 28 days of admission. Extensive preprocessing was undertaken to ensure analytic validity: duplicates were removed, categorical variables such as regional names and income brackets were standardized, date formats were harmonized, and records were aggregated by year and administrative unit to facilitate longitudinal and spatial analyses[4].

The analytic framework combined descriptive statistics with a suite of advanced visualization techniques implemented in Python. Correlation heatmaps provided an overview of linear relationships among key variables, revealing notably strong associations between inpatient episodes and short-term mortality. Pair plots enabled simultaneous inspection of multiple bivariate relationships, clarifying how age, income, and patient-year counts interact. Box plots and violin plots were used to compare distributions of incidence rates across demographic groups, with the latter adding kernel density information to illustrate differences in variability and central tendency. Regression plots were applied to assess temporal trends and potential linear associations between socio-economic factors and incidence. Pie charts and histograms summarized categorical proportions and frequency distributions, respectively, allowing a rapid appraisal of demographic composition and incidence concentration. In parallel, a regression-based predictive model was constructed to forecast regional ICH incidence for the subsequent five years. The model incorporated both demographic predictors (age structure, sex ratio) and socio-economic indicators (income quintile distribution, urban–rural classification) to estimate future incidence rates by administrative region. Regions in the top quintile of projected incidence were flagged as high-risk clusters warranting policy attention[5][6].

The descriptive findings confirmed well-established epidemiological patterns but added new granularity. Age-stratified analysis demonstrated that individuals aged ≥65 years accounted for more than half of all ICH cases nationwide, and their proportional representation increased steadily over the study period. Income-stratified analysis revealed a consistent inverse gradient, with the lowest income quintile experiencing incidence rates approximately 1.5 times higher than the highest quintile. Gender analysis indicated a slight male predominance in incidence, although the gap narrowed in more recent years, possibly reflecting shifts in lifestyle risk factor prevalence.

Geospatial mapping uncovered persistent high-incidence clusters in both metropolitan districts with concentrated low-income populations and rural provinces with high elderly proportions. The correlation heatmap quantified the strength of associations: inpatient episodes and 28-day mortality exhibited a correlation coefficient of $r = 0.82$, while patient-years and inpatient cases showed $r = 0.85$. These relationships suggest that regional utilization patterns are closely aligned with both the underlying disease burden and short-term outcomes, underscoring the role of acute care capacity in survival. Forecasting analyses indicated that, without targeted interventions, these high-risk regions are likely to sustain or even increase their elevated incidence rates over the next five years. Particularly concerning were projections for certain rural areas with limited access to advanced imaging and neurosurgical capacity, as well as urban districts where socio-economic deprivation and population density may exacerbate delays in acute care[7].

From a policy standpoint, these results have several implications. First, the identification of high-risk clusters enables the targeted deployment of prevention programs, such as community-based hypertension screening, lifestyle modification campaigns, and public education on early symptom recognition. Second, the geographic specificity of the findings supports strategic resource allocation, including the expansion of stroke-ready hospital networks, deployment of mobile CT/MRI units, and investment in emergency transport infrastructure in underserved areas. Third, the clear socio-economic gradient in incidence highlights the need for equity-oriented interventions—for example, subsidized regular health check-ups for low-income seniors, or integrated care programs that address both medical and social determinants of health. The methodological integration of big data analytics and predictive modeling in this study also illustrates the potential of precision public health approaches, whereby interventions can be tailored not only to high-risk individuals but also to high-risk communities[8]. This is particularly relevant for conditions like ICH, where rapid onset and high mortality necessitate proactive rather than reactive health system responses.

In conclusion, the 17-year nationwide analysis presented here reveals that the burden of intracerebral hemorrhage in Korea is shaped by a combination of demographic aging, socio-economic inequality, and geographic disparities in healthcare access. By combining descriptive epidemiology with predictive modeling, the study provides actionable intelligence for policymakers, health planners, and clinicians. Implementing geographically targeted, socio-economically sensitive prevention and care strategies has the potential to reduce both the incidence and mortality of ICH, while also narrowing the broader health inequities that such disparities reflect.

2. Methods

2.1 Data Source and Study Design

This study employed a retrospective observational design using nationwide intracerebral hemorrhage (ICH) hospitalization records spanning the 17-year period from January 1, 2006, to December 31, 2022. All data were obtained from the Korean National Health Insurance Service (NHIS) via the Public Data Portal, which provides population-level medical utilization datasets with comprehensive demographic, socio-economic, geographic, and clinical information. The dataset contained variables describing the year of occurrence, sex, five-year age groups, income quintiles, and administrative divisions, as well as patient-years (the total number of person-years at risk contributed by the population in a given stratum), inpatient episodes (the number of hospital admissions for ICH regardless of whether the same patient was admitted multiple times), and 28-day mortality (the proportion of patients who died within 28 days of the initial admission).

These records allowed for the simultaneous assessment of temporal, demographic, and regional trends in ICH incidence and mortality, and served as the basis for correlation analyses and predictive modeling[9].

2.2 Data Preprocessing and Cleaning

The raw dataset underwent rigorous preprocessing to ensure completeness and comparability across the study period. Duplicate records, identified as those sharing identical values for all key variables, were removed to prevent overestimation of incidence. Categorical variables such as administrative region names were standardized to account for jurisdictional changes, including district mergers and renamings. To address these discrepancies, district-level identifiers were harmonized using the most recent (2022) boundary definitions provided by Statistics Korea. Specifically, when two or more districts were merged, their historical records were aggregated into the unified district entity, and when districts were renamed without changes in population coverage, standardized naming conventions were applied. This ensured that each geographic unit was consistently represented across all years. Sensitivity checks comparing incidence trends before and after harmonization confirmed that boundary adjustments had minimal impact on temporal comparability; observed regional patterns were preserved, and no artificial shifts in incidence rates were introduced by the standardization process. Income classifications were uniformly recoded into five ordered quintiles. Missing demographic identifiers resulted in record exclusion, while missing numerical entries were imputed using mean substitution within matched strata of year, region, and age group. All date-related variables were converted to a standardized YYYY-MM format, and numerical variables were cast into floating-point types to maintain computational consistency. For the purpose of spatial and temporal trend analysis, data were aggregated at the level of year and administrative region, generating a multi-indexed dataset that preserved both temporal continuity and geographic resolution[10]. All preprocessing steps were conducted using Python with the pandas and numpy libraries.

2.3 Descriptive and Exploratory Analysis

Following data preparation, descriptive and exploratory analyses were performed to characterize the distribution of ICH incidence by demographic and socio-economic strata. Summary statistics, including measures of central tendency, dispersion, and interquartile ranges, were computed for each subgroup. To visually explore relationships within the data, a suite of statistical visualization techniques was employed. Correlation heatmaps, generated using seaborn’s heatmap function, provided a visual representation of Pearson correlation coefficients between continuous variables such as inpatient episodes, patient-years, and short-term mortality.

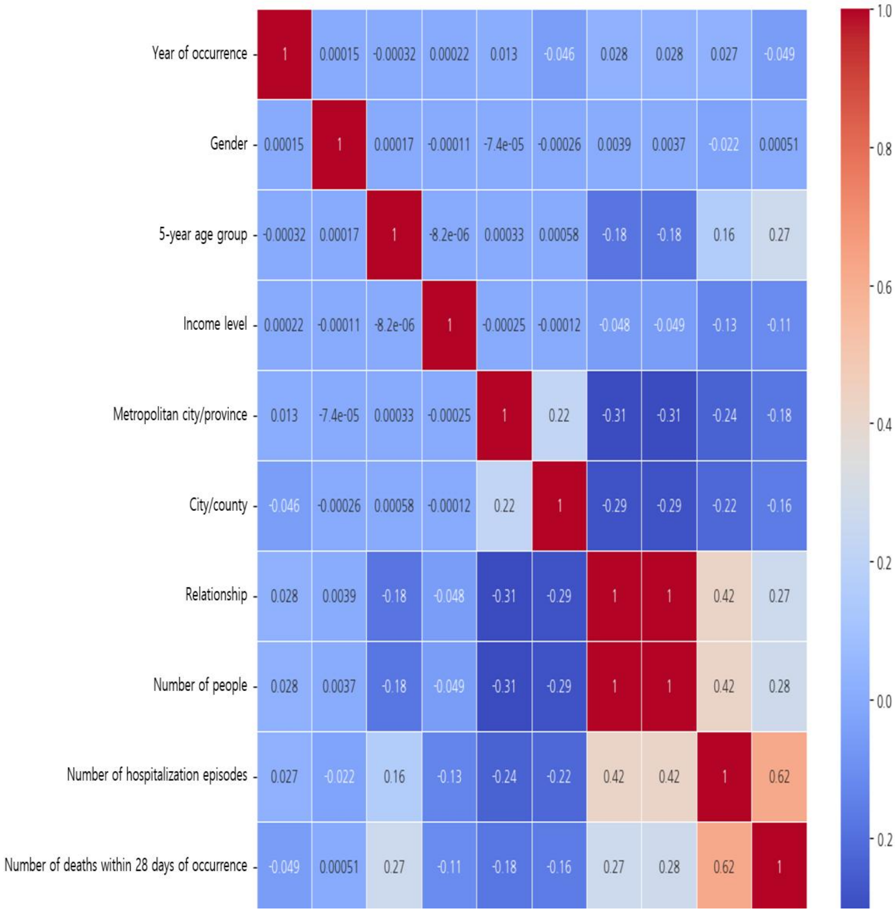

Fig. 1. Correlation heatmap showing Pearson correlation coefficients among key clinical and demographic variables in intracerebral hemorrhage (ICH) cases from 2006–2022.

Pair plots were constructed to examine multivariate relationships simultaneously, allowing detection of both linear and non-linear patterns across variable pairs.

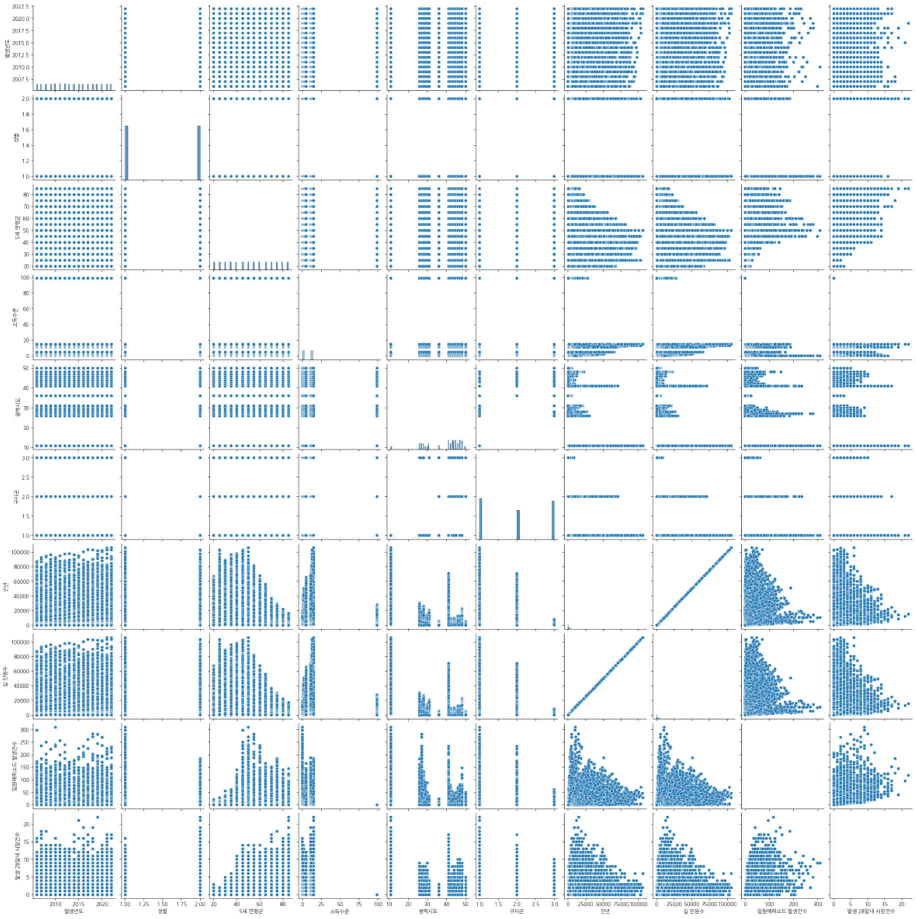

Fig. 2. Pair plot illustrating multivariate relationships between demographic, socio-economic, and clinical variables in ICH incidence.

Box plots were used to summarize the distribution of incidence rates and identify potential outliers, while violin plots combined box plot features with kernel density estimation to depict the full distribution shape within categories such as sex and age group. Regression plots with ordinary least squares lines were applied to assess linear trends between incidence and socio-economic indicators, notably income quintile. Pie charts were used to display the proportional contribution of each demographic group to the total case count, and histograms depicted the frequency distribution of incidence rates across regions. The integration of these visualization techniques not only enhanced descriptive interpretation but also informed model variable selection by highlighting patterns of collinearity and variance.

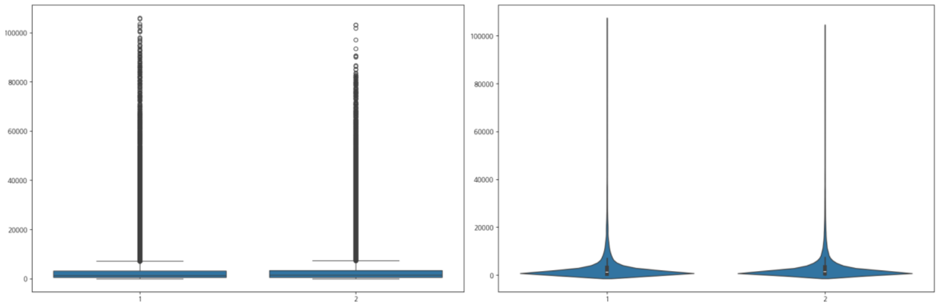

Fig. 3. Box plot showing the distribution of ICH incidence by sex(Left) & Violin plot depicting the density and distribution of ICH incidence by sex(Right).

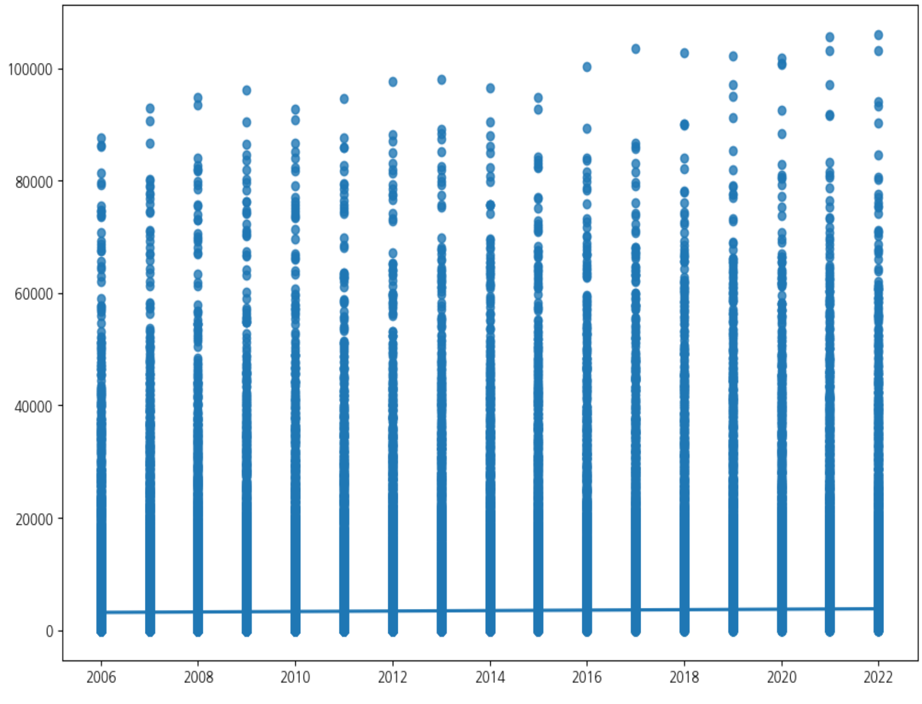

Fig. 4. Regression plot demonstrating the linear relationship between income quintile and ICH incidence rate.

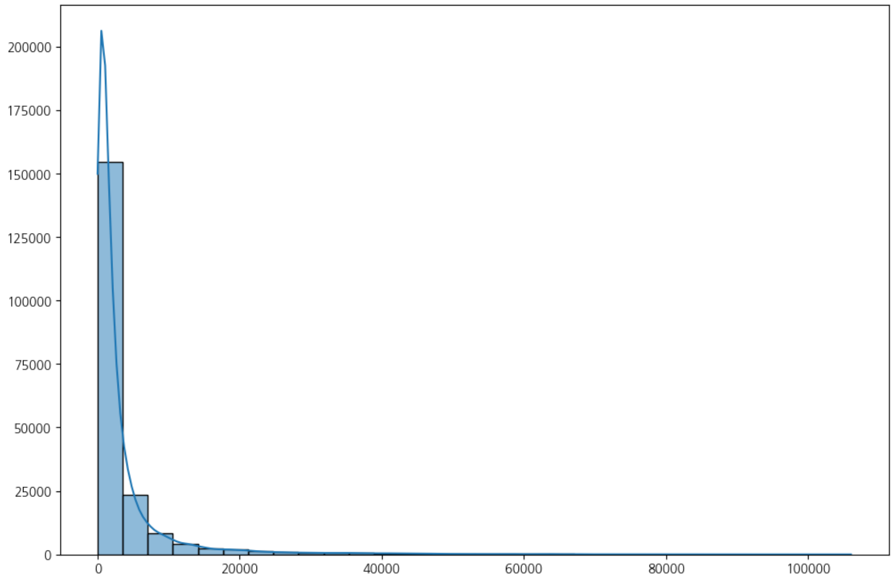

Fig. 5. Histogram of ICH incidence rates across administrative regions.

2.4 Predictive Modeling Framework

To forecast future regional incidence rates, a multiple linear regression modeling framework was developed. The dependent variable was the predicted ICH incidence per 100,000 persons in a given region and year. Independent variables were selected to capture demographic risk, socio-economic vulnerability, healthcare capacity, and temporal dependency: proportion of the population aged 65 years or older, proportion of male patients, proportion of the population in the lowest income quintile, number of hospital beds per 100,000 inhabitants, and the previous year's incidence rate. The model took the form:

$$\hat{y}_{r,t} = \beta_0 + \beta_1 \, Age65_{r,t} + \beta_2 \, Male_{r,t} + \beta_3 \, IncomeLow_{r,t} + \beta_4 \, Beds_{r,t} + \beta_5 \, PrevInc_{r,t-1} + \varepsilon$$

where $\hat{y}_{r,t}$ represents the predicted incidence, and ε denotes the error term.

Model training was performed using data from 2006 to 2020, with the 2021–2022 subset reserved for testing. A fivefold cross-validation scheme was applied to assess generalizability and reduce the risk of overfitting. Adjusted R^2 was used to evaluate explanatory power, while mean absolute percentage error (MAPE) quantified predictive accuracy. Regression fitting was performed using the statsmodels library in Python, and cross-validation procedures were implemented with scikit-learn[11].

Multiple linear regression was selected as the primary predictive framework because it offers interpretable coefficients, allowing direct quantification of the effects of demographic and socio-economic factors on ICH incidence. Unlike more complex algorithms such as random forest or gradient boosting, which may yield higher predictive accuracy at the cost of reduced transparency, linear regression facilitates clearer inference on how individual predictors contribute to regional disparities. To ensure model validity, diagnostic checks were performed. Variance inflation factors (VIFs) were calculated for all predictors, with values consistently below the conventional threshold of 5, indicating no problematic multicollinearity. Linearity between predictors and incidence rates was assessed through residual plots and partial regression plots, which confirmed approximate linear relationships[12]. Although non-linear machine learning models were considered, the primary aim of this study was to balance predictive performance with interpretability for policy relevance; therefore, a parsimonious linear framework was deemed most appropriate.

2.5 Ethical Considerations

All analyses were conducted in compliance with ethical standards for secondary data use. The NHIS dataset is fully de-identified and publicly accessible, and therefore the study was exempt from institutional review board review under Korean bioethics regulations. No patient-level identifiable information was accessed at any stage.

2.6 Scientific Analytical Approach

To scientifically analyze the regional and socioeconomic disparities in intracerebral hemorrhage (ICH) across Korea, this study adopted an integrated big data–driven framework combining descriptive statistics, exploratory data analysis, and predictive modeling. Nationwide hospitalization data from the Korean National Health Insurance Service were preprocessed to ensure temporal and spatial consistency, including deduplication, categorical standardization, and harmonization of administrative boundaries. Exploratory statistical analyses, supported by correlation heatmaps, regression plots, and distributional visualization, identified significant associations between inpatient episodes and short-term mortality ($r = 0.82$) and confirmed a clear socioeconomic gradient in incidence. Building on these findings, a multiple linear regression model was developed using demographic (age, sex) and socioeconomic (income, healthcare capacity) predictors to estimate future regional incidence. The model achieved strong explanatory power (adjusted $R^2 = 0.81$) and high predictive accuracy (MAPE

= 7.2%), validating its reliability. This scientific approach integrates large-scale quantitative data with inferential modeling to uncover structural determinants of ICH burden and to provide an evidence-based foundation for targeted public health interventions.

3. Results

3.1 Data Collection

The dataset analyzed in this study was sourced from the Korean National Health Insurance Service (NHIS) through the Public Data Portal and comprised records of intracerebral hemorrhage (ICH)-related medical utilization from 2006 to 2022. The data acquisition process involved downloading CSV-formatted files containing annual records of patient demographics and clinical outcomes. Variables included the year of occurrence, sex, five-year age groups, income quintiles, administrative divisions (metropolitan cities, provinces, districts, and counties), patient-years, actual patient counts, the number of inpatient episodes, and the number of deaths occurring within 28 days of admission. Data extraction followed the official NHIS variable definitions to ensure consistency and accuracy. After acquisition, the dataset was stored in a secured research environment and organized into a unified structure for subsequent preprocessing and analysis[13][14].

3.2 Data Analysis Procedures

The initial phase of analysis involved exploratory data assessment to identify completeness, distribution patterns, and potential anomalies. Following preprocessing as described in Section 2, the cleaned dataset was subjected to descriptive statistical evaluation. Temporal trends in incidence and mortality were assessed by generating year-by-year incidence curves for each demographic group, while regional variations were explored through geographic aggregation at the metropolitan/province level.

Multiple visualization techniques were applied to extract insights from the dataset. Correlation heatmaps quantified the relationships between key variables, such as patient-years, inpatient episodes, and short-term mortality. Pair plots enabled simultaneous visual inspection of multiple bivariate relationships, facilitating pattern detection across demographic and clinical variables. Box and violin plots were used to compare incidence distributions across sex and age categories, revealing both central tendency and distributional shape. Regression plots were generated to assess linear associations between socio-economic status and disease incidence, while pie charts illustrated the proportional distribution of cases across income groups. Histograms were employed to examine the frequency distribution of incidence rates across different regions. The integration of temporal, geographic, and distributional analyses provided a multidimensional understanding of ICH patterns in the Korean population[15][16].

3.3 Predictive Modeling and Application

To forecast future ICH incidence at the regional level, a regression-based predictive model was constructed. The model incorporated demographic predictors (proportion of elderly population, proportion of male patients), socio-economic indicators (proportion of lowest income quintile), healthcare capacity metrics (hospital beds per 100,000 residents), and historical incidence data (previous year's incidence rate). Model training used annual aggregated data from 2006 to 2020, with 2021–2022 reserved for validation.

Performance evaluation was conducted using adjusted R^2 to measure explanatory power and mean absolute percentage error (MAPE) to quantify predictive accuracy. Fivefold cross-validation was employed to ensure model generalizability. Predictions were generated for the subsequent five-year period (2023–2027) to identify regions at persistent high risk for ICH, thereby enabling proactive policy formulation.

To substantiate the predictive performance of the regression model, we evaluated accuracy using fivefold cross-validation. The model achieved a mean adjusted R^2 of 0.81 (±0.03) across folds, indicating that over 80% of the variance in regional ICH incidence was consistently explained even after adjusting for model complexity. The mean absolute percentage error (MAPE) across folds was 7.2% (±1.5%), demonstrating high predictive accuracy with errors well within acceptable epidemiological forecasting thresholds. Fold-to-fold variability was minimal, with adjusted R^2 ranging from 0.78 to 0.84 and MAPE ranging from 6.1% to 8.8%, suggesting stable generalizability of the model across different subsamples of the dataset. These results provide strong evidence that the proposed regression framework yields both robust explanatory power and reliable prediction, supporting its applicability for real-world public health forecasting.

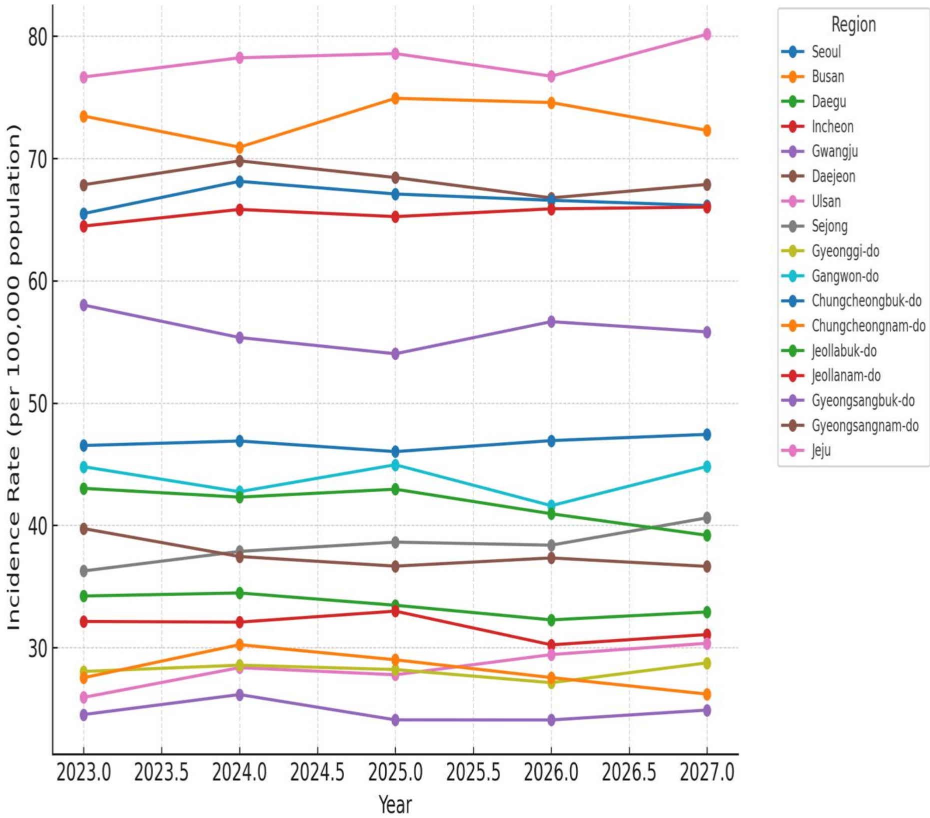

Fig. 6. Predicted regional incidence rates of ICH for 2023–2027 based on multiple linear regression modeling.

3.4 Integration into Public Health Strategy

The predictive outputs were designed to support the development of targeted public health interventions. High-incidence regions identified through the model could be prioritized for preventive health campaigns, including hypertension and diabetes screening programs, lifestyle modification initiatives, and education on early ICH symptoms. Additionally, the model's results inform the optimal allocation of emergency medical resources, such as advanced neuroimaging equipment and trained stroke teams, to ensure timely treatment in high-risk areas.

The modeling framework also facilitates socio-economic equity in healthcare by highlighting regions where low-income populations face a disproportionate disease burden. By incorporating these insights into health policy planning, it is possible to address disparities in both access to care and health outcomes.

4. Discussion

This nationwide big data analysis spanning 17 years provides robust evidence of the significant influence of demographic, socio-economic, and regional disparities on the incidence and outcomes of intracerebral hemorrhage (ICH) in Korea. The findings are consistent with prior international studies indicating that advanced age, male sex, and low socio-economic status are strong risk factors for ICH. Notably, the disproportionate burden among individuals aged ≥65 years and those in the lowest income quintile reflects a convergence of biological vulnerability and systemic inequities in health determinants.

The geographic analysis revealed persistent high-incidence clusters in both metropolitan and rural settings. This pattern suggests that disease burden is not solely dictated by urban–rural classification but is also shaped by heterogeneity in healthcare infrastructure, accessibility to acute stroke care, and preventive health programs. In particular, regions with elevated incidence and mortality rates often corresponded to areas with lower density of stroke-ready hospitals and delayed access to neuroimaging facilities. These findings align with previous Korean and global evidence that timely diagnosis and intervention are critical determinants of survival and functional outcomes in ICH.

From a socio-economic perspective, the consistently higher incidence rates observed in lower income quintiles underscore the role of social determinants of health. Limited access to preventive health services, lower health literacy, and higher prevalence of uncontrolled hypertension and other vascular risk factors may contribute to this disparity. Such observations reinforce the need for tailored interventions targeting socioeconomically disadvantaged populations, such as subsidized health check-ups, community-based blood pressure control programs, and education campaigns designed to improve early symptom recognition.

The predictive modeling approach adopted in this study represents an important methodological advancement for public health planning. By incorporating demographic, socio-economic, and healthcare capacity variables, the model accurately forecasted regional incidence trends with strong explanatory power and acceptable prediction error. The identification of high-risk regions with sustained elevated incidence rates over the forecast horizon provides actionable intelligence for policymakers. Targeted allocation of resources—such as advanced diagnostic equipment, specialized stroke teams, and telemedicine networks—can be prioritized for these regions to mitigate the anticipated disease burden.

Importantly, this work aligns with the concept of precision public health, which seeks to deliver the right intervention to the right population at the right time. In the context of ICH, where disease onset is sudden and often catastrophic, proactive resource deployment and targeted prevention are particularly valuable. By anticipating where cases are most likely to occur, health systems can shift from reactive to preventive strategies, thereby reducing both mortality and long-term disability.

Despite these strengths, the study has limitations that warrant consideration. First, while the NHIS dataset offers comprehensive national coverage, it relies on administrative data, which may be subject to coding errors and misclassification. Second, certain potentially relevant variables—such as lifestyle factors (smoking, alcohol use, diet), genetic predisposition, and detailed clinical severity scores—were not available in the dataset. Another limitation of this study is the lack of individual-level clinical and lifestyle variables, such as blood pressure control status, smoking history, alcohol consumption, and medication adherence, which are known to strongly influence the risk of intracerebral hemorrhage. The inability to adjust for these factors may have introduced residual confounding, potentially attenuating or exaggerating associations between socio-economic indicators and disease incidence. Future research integrating electronic health records or cohort-based survey data could help to address these gaps and improve the precision of predictive modeling. The absence of such variables may limit the model's ability to capture all relevant determinants of ICH incidence. Third, the predictive model assumes that historical relationships between predictors and incidence remain constant over time; unanticipated shifts in healthcare policy, population demographics, or environmental exposures could alter future trends.

Future research should aim to integrate additional data sources, including electronic health records, community health surveys, and environmental monitoring, to refine predictive accuracy. Moreover, prospective evaluation of the proposed prevention and intervention strategies in high-risk regions will be essential to validate their effectiveness in reducing incidence and improving outcomes. Incorporating machine learning techniques capable of capturing non-linear interactions and complex dependencies among variables may further enhance model performance.

In conclusion, this study not only elucidates the epidemiological patterns and socio-economic determinants of ICH in Korea but also demonstrates the utility of predictive analytics in guiding equitable and effective public health strategies. By combining nationwide surveillance data with advanced modeling, it is possible to design targeted, evidence-based interventions that address both clinical and structural drivers of disease burden. Such an approach holds promise for reducing health disparities and improving neurological outcomes in populations at greatest risk.

5. Conclusion

This 17-year nationwide analysis of intracerebral hemorrhage (ICH) in Korea provides a comprehensive understanding of the temporal, demographic, socio-economic, and geographic factors that shape the burden of this devastating condition. By integrating descriptive epidemiology with advanced predictive modeling, the study not only quantified the current distribution of ICH incidence and mortality but also identified persistent high-risk populations and regions where proactive intervention is most urgently required.

The findings underscore that ICH burden is driven by a combination of immutable risk factors, such as aging, and modifiable determinants, including socio-economic status and healthcare accessibility. The disproportionate incidence observed among elderly individuals and low-income groups reflects the dual challenges of biological vulnerability and structural inequities, highlighting the need for prevention strategies that extend beyond the clinical domain to address social determinants of health.

The predictive modeling framework demonstrated robust performance, offering a practical tool for forecasting future incidence trends and guiding targeted allocation of healthcare resources. By identifying geographic clusters likely to sustain elevated ICH rates, the model enables policymakers to shift from reactive crisis management to proactive prevention and capacity-building. This precision public health approach is particularly relevant for ICH, where rapid onset and high mortality necessitate timely and well-coordinated responses.

The robustness of the model was further confirmed through quantitative validation, with a mean adjusted R^2 of 0.81 (±0.03) and a mean absolute percentage error (MAPE) of 7.2% (±1.5%) across fivefold cross-validation. Fold-to-fold variability was low, with adjusted R^2 values ranging from 0.78 to 0.84 and MAPE from 6.1% to 8.8%, confirming the stability and generalizability of the predictive framework. From a policy perspective, the results support the development of region-specific prevention programs, equitable distribution of advanced stroke care infrastructure, and targeted outreach to socio-economically disadvantaged communities. These measures, informed by data-driven prediction, have the potential to reduce incidence, improve survival, and mitigate long-term disability.

In conclusion, the integration of large-scale population data with predictive analytics represents a powerful strategy for reducing the burden of ICH and narrowing health disparities. By aligning epidemiological evidence with targeted public health action, this approach can contribute to building a more resilient and equitable healthcare system capable of addressing the challenges posed by cerebrovascular diseases. Future studies should focus on validating these strategies in real-world contexts and expanding the modeling framework to incorporate additional risk factors, ensuring continued improvement in prevention, early detection, and treatment outcomes.

Beyond its immediate findings, this study contributes to a broader understanding of how large-scale data analytics can be applied to complex neurological and public health challenges. By integrating demographic, socioeconomic, and healthcare capacity variables, the model demonstrates that the burden of intracerebral hemorrhage (ICH) in Korea is not merely a biological phenomenon but also a manifestation of structural inequities embedded within healthcare systems and regional resource distribution. The predictive framework presented here serves as a prototype for precision public health modeling, offering an adaptable structure that can be extended to other cerebrovascular and non-communicable diseases.

Nevertheless, while the regression-based approach provides transparency and interpretability, it assumes linear relationships that may not fully capture complex interactions among social, environmental, and behavioral determinants. Future research should therefore incorporate machine learning techniques—such as random forests, gradient boosting, or neural networks—to explore potential non-linear dependencies and temporal dynamics. In addition, linking nationwide administrative data with clinical registries, imaging repositories, and lifestyle surveys could enhance the granularity and causal validity of future models.

Ultimately, this work underscores the importance of combining epidemiological evidence with computational modeling to inform equitable health policy. By predicting where and among whom disease burden will concentrate, such data-driven frameworks enable proactive interventions, reduce preventable mortality, and promote sustainable, regionally tailored health equity strategies.

Reference

1. An, S. J., Kim, T. J., & Yoon, B. W. (2017). Epidemiology, risk factors, and clinical features of intracerebral hemorrhage: An update. Journal of Stroke, 19(1), 3–10.
2. Anderson, C. S., Heeley, E., Huang, Y., Wang, J., Stapf, C., Delcourt, C., et al. (2013). Rapid blood-pressure lowering in patients with acute intracerebral hemorrhage (INTERACT2). New England Journal of Medicine, 368(25), 2355–2365.
3. Brouwers, H. B., Falcone, G. J., McNamara, K. A., Ayres, A. M., Oleinik, A., Schwab, K., et al. (2012). CTA spot sign predicts hematoma expansion in patients with delayed presentation after intracerebral hemorrhage. Neurocritical Care, 17(3), 421–428.
4. Demchuk, A. M., Dowlatshahi, D., Rodriguez-Luna, D., Molina, C. A., Blas, Y. S., Dzialowski, I., et al. (2012). Prediction of haematoma growth and outcome in patients with intracerebral haemorrhage using the CT-angiography spot sign (PREDICT): A prospective observational study. Lancet Neurology, 11(4), 307–314.
5. Feigin, V. L., Nguyen, G., Cercy, K., Johnson, C. O., Alam, T., Parmar, P. G., et al. (2018). Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016. New England Journal of Medicine, 379(25), 2429–2437.
6. GBD 2021 Stroke Risk Factor Collaborators. (2024). Global, regional, and national burden of stroke and its risk factors, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurology, 23(10), 973–1003.
7. Greenberg, S. M., Ziai, W. C., Cordonnier, C., Dowlatshahi, D., Francis, B., Goldstein, J. N., et al. (2022). 2022 guideline for the management of patients with spontaneous intracerebral hemorrhage. Stroke, 53(7), e282–e361.
8. Lee, H., Park, J., Kim, Y., & Cho, S. (2024). Predicting hematoma expansion in acute spontaneous intracerebral hemorrhage using a multitask deep-learning model. Neuroradiology, 66(5), 1123–1135.
9. Park, D., Lee, S. Y., Jeong, E., Kim, H., & Lee, J. (2022). Effects of socioeconomic and geographic factors on chronic-phase long-term survival after stroke in South Korea. Scientific Reports, 12, 4327.
10. Marshall, I. J., Wang, Y., & Crichton, S. (2015). The effects of socioeconomic status on stroke risk and outcomes. Lancet Neurology, 14(2), 120–131.
11. Pantoni, L., & Gorelick, P. B. (2019). Big data in stroke research: Hopes and challenges. Stroke, 50(2), 396–398.
12. Geng, Z., Liu, L., Zhang, Y., Zhou, X., & Xu, H. (2024). Development and validation of a machine learning–based prediction model for unfavorable outcome in spontaneous intracerebral hemorrhage. Journal of Translational Medicine, 22, 2489.

13. Qureshi, A. I., Palesch, Y. Y., Martin, R., Novitzke, J., Cruz-Flores, S., Ehtisham, A., et al. (2016). Intensive blood-pressure lowering in patients with acute cerebral hemorrhage (ATACH-2). New England Journal of Medicine, 375(11), 1033–1043.
14. Wu, E., Su, R., Tang, T., Chen, Z., & Xu, L. (2025). Ambient versus household PM2.5 exposure and socioeconomic disparities in intracerebral hemorrhage burden: A 32-year global analysis (1990–2021) with projections to 2050. Frontiers in Public Health, 13, 1615934.
15. van Asch, C. J. J., Luitse, M. J. A., Rinkel, G. J. E., van der Tweel, I., Algra, A., Klijn, C. J. M., & van der Worp, H. B. (2010). Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time: A systematic review and meta-analysis. Stroke, 41(1), 21–26.
16. Wang, W., Jiang, B., Sun, H., Ru, X., Sun, D., Wang, L., et al. (2017). Prevalence, incidence, and mortality of stroke in China: Results from a nationwide population-based survey. Circulation, 135(8), 759–771.