RWD96: Estimating Healthcare Costs Associated with Cardio- and Cerebrovascular Events Using Different Statistical Models – Implications from a Large-Scale Administrative Database

Koki Idehara[1], Fei Zhao[1], Seok-Won Kim[1]
1. IQVIA Solutions Japan G.K., Tokyo, Japan

CONFLICT OF INTEREST

KI, FZ, and SWK are employees of IQVIA Solutions Japan G.K., which fully funds this work.

OBJECTIVE

How can disease- and event-related healthcare costs be estimated using real-world data?

As cost-effectiveness analysis gains broader application in healthcare decision-making, the estimation of healthcare costs has become increasingly important. Nevertheless, challenges remain in capturing relevant costs robustly and accurately while ensuring reliability.

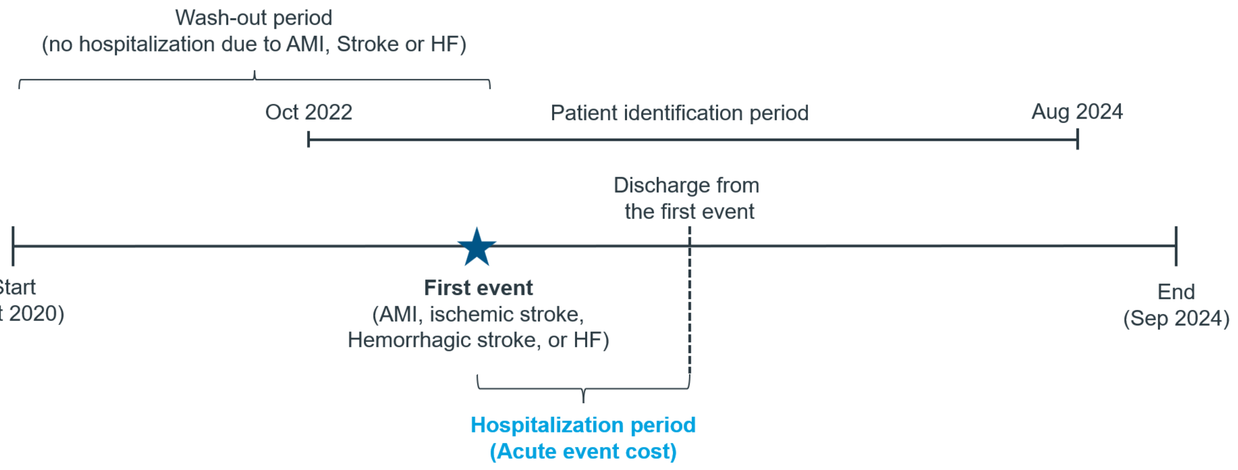
This study evaluated various regression models for estimating costs related to cardio- and cerebrovascular events, using a large-scale administrative database covering all age groups—the *IQVIA Claims Plus2 Database*.

METHODS

Overview

This study investigated model performance by assessing prediction accuracies between actual (observed) and predicted healthcare costs based on various regression models for the acute event and their follow-up costs associated with cardio- and cerebrovascular events.

Study Populations	Patients having at least one record of acute cardio- or cerebrovascular events: • Acute myocardial infarction (AMI) (see definition in Shima et al. [1]) • Ischemic stroke (IST) (see definition in Shima et al. [1]) • Hemorrhage stroke (HST) (see definition in Shima et al. [1]) • Heart failure (HF) (see definition in Nakai et al. [2])
Outcomes	Healthcare costs related to the first hospitalization (per person per episode) (cost year: as of June 2024; 1USD = 147JPY)
Study Period	Between October 2020 and September 2024
Data Source	IQVIA Claims Plus2 (an insurer-based administrative database covering all ages)
Validation Strategy	One-third of each population was randomly sampled for evaluation, while the models were developed using the remaining two-thirds.



Predictive Performance Across Different Regression Models

According to Austin et al. [3] and Malehi et al. [4], the following metrics were applied:

	Performance Metric	Formula*	Description
1	Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{i=n}([cost]_i-\widehat{[cost]}_i)^2}$	Square-root error of actual and predicted costs. Emphasized a larger difference.
2	Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{i=n}\lvert[cost]_i-\widehat{[cost]}_i\rvert$	Similar to #1 but measuring absolute error. Less sensitive to outliers than #1.
3	Mean Relative Squared Error (MRSE)	$\frac{1}{n}\sum_{i=1}^{i=n}(\frac{\widehat{[cost]}_i-[cost]_k}{[cost]_i+1})^2$	Mean squared difference between predicted and actual costs. MRSE is particularly sensitive to errors in cases whether the actual value is small.
4	Bias	$\frac{1}{n}\sum_{i=1}^{i=n}\widehat{[cost]}_i-\frac{1}{n}\sum_{i=1}^{i=n}[cost]_i$	Difference between mean predicted and actual costs. Providing direction and extent of systematic error.

*n: Number of patients; $[cost]_i$: Actual (observed) cost for i's patient; $\widehat{[cost]}_i$: Predicted cost for i's patient

Regression Models

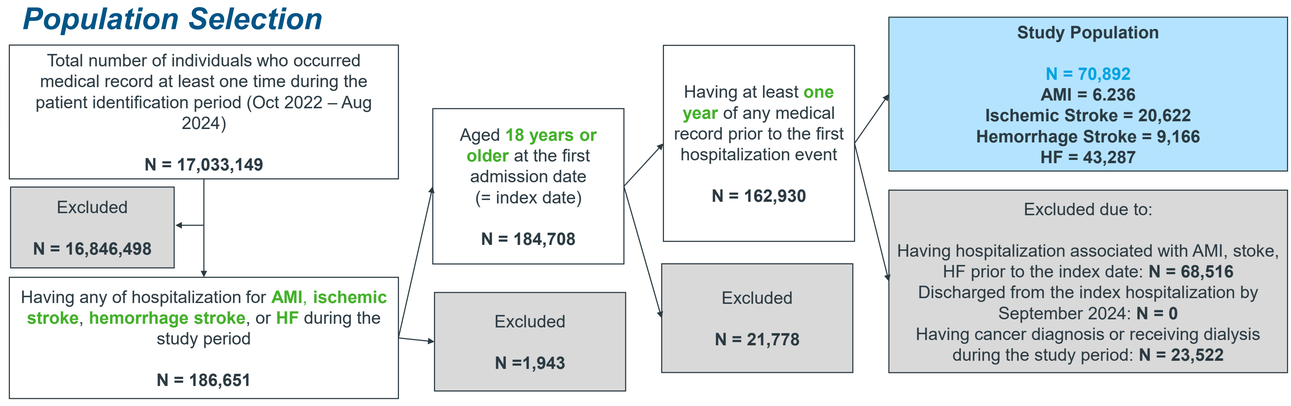
Multivariable regression models were constructed with age group (18 – 64 yrs / 65 – 74 yrs / 75 – 84 yrs / 85 yrs and above), sex, Charlson Comorbidity Index (CCI)[5,6], and prevalent comorbidities among the study populations (i.e., hypertension [HT], diabetes [DM], prior ischemic heart disease [IHD], transient ischemic attack [TIA], atrial fibrillation [AF], peripheral arterial disease [PAD], and kidney disease [KD]).

The regression models of interests are as follows:

Abb	Regression Model	Description
LN	Linear (ordinary least square)	Predicting healthcare costs by modeling the linear relationship between covariates and the outcome. Sensitive to outliers [3,7].
LL	Linear with log-transformed cost (smearing estimate proposed by Duan [8])	Similar to LN but predicts costs using a log-transformed cost variable, resulting in more normalized distribution. However, misspecification of the outcome distribution can reduce the model performance [9]. In this study **'smearing estimate'** proposed by Duan [8] was applied (i.e., $\widehat{[cost]}_i = \exp(\widehat{[\log\_cost]}_i) \times \frac{1}{n}\sum_{i=1}^{i=n}\exp(\varepsilon_i)$).
GM	Gamma (with log-link)	Applying generalized linear models by assuming the outcome distribution follows gamma, Poisson, or negative binomial distributions. Due to dealing with skewness, log-link functions are generally used [3,4,7,9].
PO	Poisson (with log-link)	
NB	Negative binomial (with log-link)	
ME	Median	Predicting median healthcare costs by estimating the conditional median of the cost distribution. Robust to outliers and skewed data [3].
CO	Cox proportional hazards	Healthcare costs are modeled based on the relationship between covariates and the hazard of attaining final cost, using a framework where cost replaces survival time. Censoring can be taken into account. Not requiring distributional assumptions [3,4,10].

RESULTS

Population Selection

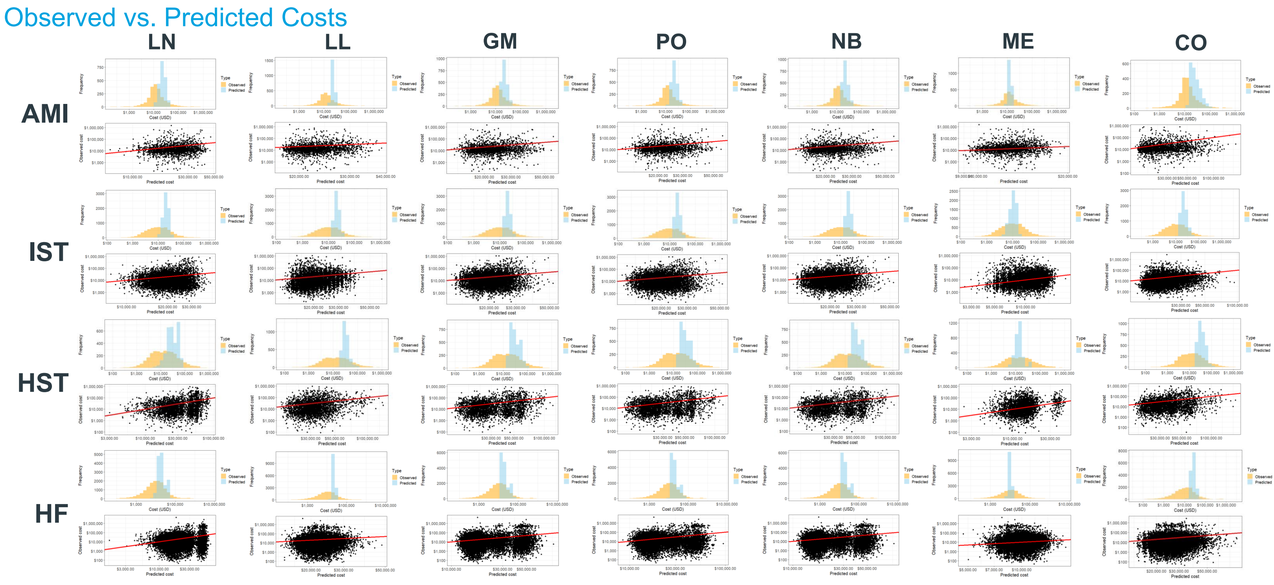


Patient Characteristics

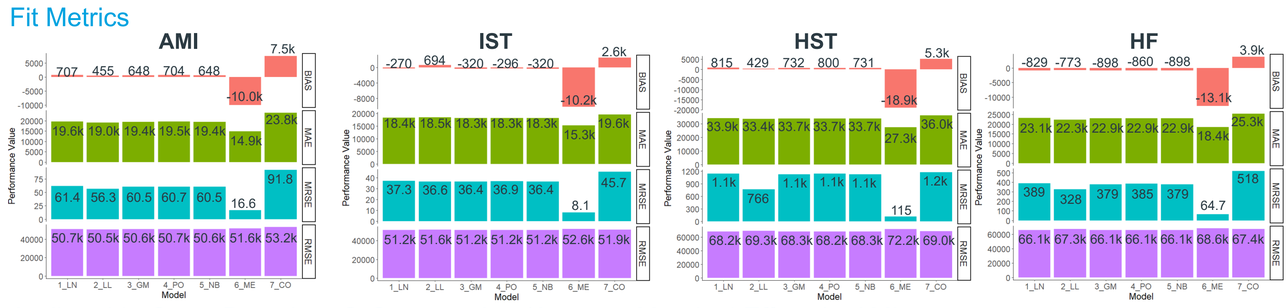
Variable	AMI (N = 6,236)	Ischemic stroke (IST) (N = 20,622)	Hemorrhage stroke (HST) (N = 9,166)	HF (N = 43,287)
Age group (n, %)				
18 – 64 years	2,255 (36.2%)	4,086 (19.8%)	2,998 (32.7%)	6,495 (15.0%)
65 – 74 years	1,164 (18.7%)	3,293 (16.0%)	1,402 (15.3%)	4,959 (11.5%)
75 – 84 years	1,635 (26.2%)	6,729 (32.6%)	2,514 (27.4%)	12,210 (28.2%)
85 years or above	1,182 (19.0%)	6,514 (31.6%)	2,252 (24,6%)	19,624 (45.3%)
Male sex (n, %)	4,485 (71.9%)	10,846 (52.6%)	4,665 (50.9%)	20,509 (47.4%)
Baseline CCI score	4.3 ± 2.0	3.9 ± 2.0	3.5 ± 2.0	4.1 ± 2.1
Hypertension (n, %)	5,758 (92.3%)	18,050 (87.5%)	8,062 (88.0%)	38,337 (88.6%)
Diabetes (n, %)	3,866 (62.0%)	10,283 (49.9%)	3,547 (38.7%)	24,012 (55.5%)
Prior IHD (n, %)	4,861 (78.0%)	5,929 (28.8%)	1,740 (19.0%)	20,978 (48.5%)
TIA (n, %)	95 (1.5%)	1,154 (5.6%)	499 (5.4%)	903 (2.1%)
AF (n, %)	1,247 (20.0%)	5,634 (27.3%)	1,153 (126%)	17,184 (39.7%)
Kidney disease (n, %)	845 (13.6%)	2,622 (12.7%)	882 (9.6%)	10,365 (239%)
PAD (n, %)	631 (10.1%)	2,017 (9.8%)	564 (6.2%)	5,700 (13.2%)
Length of event (days)	19.2± 31.5	32.0 ±37.1	36.6 ± 44.7	30.5 ± 41.0
Death during the event	369 (5.9%)	906 (4.4%)	845 (9.2%)	3,654 (8.4%)
Event cost (per episode)	$23,151 ± $57,111 (¥3,403,180±¥8,395,293)	$20,786 ± $46,602 (¥3,055,491±¥6,850,555)	$34,104 ± $70,832 (¥5,013,220±¥10,412,243)	$22,562 ± $68,469 (¥3,316,595±¥10,064,969)
Baseline cost (per year)	$3,543 ±$17,001 (¥520,777 ± ¥2,513,806)	$5,218 ± $24,734 (¥767,056±¥3,635,930)	$6,245 ± $28,002 (¥917,944±¥4,116,287)	$9,283 ± $577,971 (¥1,364,598±¥84,961,769)

Model Performance

Observed vs. Predicted Costs



Fit Metrics



Attributable Costs Based on G-Computation (Average Marginal Effects) - AMI

Variable	LN			LL (smearing estimate)			GM		
	Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI
β_0	$23,465	$14,675	$32,254	$23,392	$19,977	$27,464	$32,280	-	-
log([cost])	$135	-$712	$982	-$496	-$769	-$238	-$7	-	-
Age group (ref: 18 – 64 yrs)									
65 – 74 yrs	-$3,917	-$9,219	$1,384	-$2,817	-$9,055	$2,659	-$3,142	-	-
75 – 84 yrs	-$7,729	-$12,727	-$2,730	-$5,661	-$10,474	-$1,237	-$6,224	-	-
85 yrs or above	-$10,079	-$15,929	-$4,230	-$7,246	-$12,287	-$2,546	-$8,233	-	-
Female sex	-$2,194	-$6,615	$2,227	-$4,109	-$7,288	-$720	-$1,960	-	-
Baseline CCI score	$1,818	$606	$3,030	$1,193	$695	$1,715	$1,618	-	-
Baseline comorbidity									
HT	-$732	-$7,848	$6,384	-$1,477	-$7,840	$3,597	-$1,237	-	-
DM	-$2,274	-$6,501	$1,953	-$1,361	-$6,028	$2,461	-$2,369	-	-
Prior IHD	-$4,795	-$9,303	-$287	-$3,693	-$9,020	$680	-$4,745	-	-
TIA	-$4,132	-$18,248	$9,983	-$3,822	-$8,852	$3,449	-$3,139	-	-
AF	$11,508	$6,856	$16,161	$10,327	$5,102	$15,812	$11,051	-	-
PAD	$1,322	-$4,965	$7,608	$1,594	-$3,192	$7,368	$1,403	-	-
KD	-$571	-$6,650	$5,508	-$158	-$4,482	$4,494	$163	-	-

DISCUSSION

- Linear regression can be used as a *reference*, as it yields comparable performance to linear regression with log-transformed cost and generalized regressions, while remaining a simpler additive model
- Median regression showed better MRSE and MAE than other models but not feasible to predict mean cost (larger negative bias and larger root-mean squared error)
- Cox regression considered censoring (i.e., death during event) cases, resulting in larger positive bias and worse model fits with observed data, since the model accounted for the censored cases. The model should be considered when dealing with data having frequent censoring during the observation period.

References

[1] Shima D, Ii Y, Higa S, et al. Validation of novel identification algorithms for major adverse cardiovascular events in a Japanese claims database. The Journal of Clinical Hypertension. 2021;23(3):646-55.
[2] Nakai M, Iwanaga Y, Sumita Y, et al. Associations among cardiovascular and cerebrovascular diseases: Analysis of the nationwide claims-based JROAD-DPC dataset. PLOS ONE. 2022;17(3):e0264390.
[3] Austin PC, Ghali WA, Tu JV. A comparison of several regression models for analysing cost of CABG surgery. Stat Med. 2003;22(17):2799-815.
[4] Malehi AS, Pourmotahari F, Angali KA. Statistical models for the analysis of skewed healthcare cost data: a simulation study. Health Econ Rev. 2015;5:11.
[5] Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373-83.
[6] Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43(11):1130-9.
[7] Mihaylova B, Briggs A, O'Hagan A, et al. Review of statistical methods for analysing healthcare resources and costs. Health Economics. 2011;20(8):897-916.
[8] Duan N. Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association. 1983;78(383):605-10.
[9] Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ. 2001;20(4):461-94.
[10] MeWhinney S, Brown ER, Malcolm J, et al. Identification of risk factors for increased cost, charges, and length of stay for cardiac patients. Ann Thorac Surg. 2000;70(3):702-10.
[11] Bang H, Tsiatis AA. Estimating medical costs with censored data. Biometrika. 2000;87(2):329-43.

Contact: koki.idehara@iqvia.com

ISPOR Real-World Evidence Summit 2025, 28–30 September 2025, in Tokyo, Japan