Abstract Type : Oral presentation
Abstract Submission No.: A-0325
Abstract Topic : Acute Kidney Injury

Privacy-preserving synthetic data enhances PO-AKI prediction in data-scarce scenarios

Soie Kwon[1], Eunbyeol Cho, Minjae Lee, Edward Choi, Hajeong Lee[2]
[1]Department of Internal Medicine-Nephrology, Chung-Ang University Hospital, Korea, Republic of
[2]Department of Internal Medicine-Nephrology, Seoul National University Hospital, Korea, Republic of
[3]Department of Kim Jaechul Graduate School of Art of AI, KAIST, Korea, Republic of

Objectives : Despite the growing use of artificial intelligence, data availability, and privacy concerns limit its clinical application. This study aimed to develop a synthetic model as a promising solution to address these, enabling the prediction of postoperative acute kidney injury (PO-AKI) prediction even with a relatively small real-world dataset.
Methods : We developed a synthetic model to generate virtual patient data, incorporating comorbidities, laboratory results, medication history, surgical details, and PO-AKI occurrence in patients underwent non-cardiac major surgeries. The model was built on the BERT architecture and trained using real-world data from data-rich hospitals. Privacy risks were evaluated through Membership and Attribute Inference Attacks (MIA and AIA). The similarity between synthetic and real-world data was statistically assessed, and its clinical utility was evaluated by examining whether augmenting data-scarce scenarios with exact matched synthetic data improved PO-AKI prediction using the CatBoost.
Results : A total of 335,687 real-world patient data were collected from six tertiary hospitals, including 275,727 from 3 data-rich and 59,960 from 3 data-scarce hospitals. The similarity between the real-world data from the data-rich hospitals, which served as the training set for the synthetic generation model, and the synthetic data from each hospital was analyzed (Table 1). At SNUH, 90.4% of variables showed no statistically significant difference between real-world and synthetic data, compared to 89.0% at SNUBH and 94.4% at AMC. The MIA and AIA confirmed the privacy protection of synthesized data. The clinical utility of synthetic data in PO-AKI prediction was evaluated by augmenting real-world data-scarce cohorts (250–2,000 patients) with synthetic data. The benefit was most pronounced in smaller cohorts, peaking at 2,000–4,000 synthetic patients and plateauing beyond 16,000 (Figure 1).
Conclusions : This is the first study to apply generative AI to PO-AKI prediction. We comprehensively demonstrate its clinical utility in data-scarce scenarios by enhancing prediction performance through synthetic data augmentation.

2025_KSN_초록_Table1.png

Table 1. Similarity comparison of representative variables between real-world and synthetic data

Continuous variables were presented as mean ± standard deviation (SD), and categorical variables were presented as frequencies (percentages). As continuous variables were not normally distributed, the Kolmogorov-Smirnov (KS) statistic was used, with lower KS statistic values indicating greater similarity between the distributions of the two datasets. Medication usage information is presented as used in the 90 days before surgery. For categorical variables, Jensen-Shannon Distance (JSD) was used to measure similarity, where values closer to 0 indicate greater similarity to real data

Feature	SNUH					SNUBH					AMC				
	Real-world data		Synthetic data		*p* value	Real-world data		Synthetic data		*p* value	Real-world data		Synthetic data		*p* value
	Value	Missing (%)	Value	Missing (%)		Value	Missing (%)	Value	Missing (%)		Value	Missing (%)	Value	Missing (%)	
Age (year)	58.1±15.1	0	58.9±13.3	0	0.052	57.3±15.4	0	56.9±14.5	0	0.076	57.1±14.6	0	57.6±13.4	0	0.044
Sex, Female	46.1	0	45.7	0	0.002	52.4	0	52.1	0	0.002	50.4	0	50.0	0	0.003
Height (cm)	162.3±9.2	42	162.7±9.5	41	0.040	161.7±9.4	1	160.9±9.5	0	0.032	162.3±9.2	0	162.2±9.2	0	0.031
Weight (kg)	64.2±12.2	42	64.6±12.6	41	0.028	64.2±12.2	1	63.4±12	0	0.040	63±12.1	0	63±12.0	0	0.013
SBP (mmHg)	122.6±15.9	4	122.1±15.4	3	0.047	123.4±20.1	0	121.6±19.8	0	0.046	117.7±15.9	2	117.3±16.5	2	0.032
DBP (mmHg)	75.5±11	4	75.3±10.6	3	0.036	71.7±12.9	0	70.8±12.8	0	0.051	73.6±10	2	73.3±10.2	2	0.040
Heart rate (/min)	69.1±12.3	44	68.7±11.6	42	0.050	70.3±13.1	0	69.2±12.7	0	0.0046	69.9±11.6	2	69.5±11.8	2	0.040
Operation-related factor															
Department (%)															
GS	46.9	0	53..3	0		39.5	0	44.7	0		54.0	0	59.7	0	
OBGY	4.3	0	3.4	0		5.5	0	6.3	0		12.9	0	11.4	0	
UR	8.6	0	6.3	0	0.069	9.8	0	9.5	0	0.046	7.7	0	4.4	0	0.059
NS	16.1	0	14.8	0		9.3	0	7.2	0		10.3	0	11.0	0	
OS	27.0	0	22.2	0		35.9	0	32.4	0		15.2	0	13.4	0	
Emergent OP	7.8	0.6	7.0	0.6	0.012	5.5	0	3.6	0	0.033	5.1	0	4.9	0	0.003
Comorbidity (%)															
DM	12.6	0	12.0	0	0.006	12.4	0	11.8	0	0.006	13.6	0	13.6	0	<0.001
Hypertension	23.7	0	21.7	0	0.017	28.3	0	29.0	0	0.005	27.6	0	29.2	0	0.012
Malignancy	40.3	0	48.5	0	0.059	30.0	0	30.6	0	0.005	45.0	0	42.9	0	0.015
90-days medication usage (%)															
RAS Blocker	10.0	0	11.2	0	0.013	6.8	0	5.1	0	0.026	17.3	0	18.0	0	0.006
NSAIDs	21.3	0	17.9	0	0.031	22.7	0	16.1	0	0.059	12.1	0	14.1	0	0.021
Statin	9.4	0	8.6	0	0.010	6.1	0	4.7	0	0.023	13.7	0	13.2	0	0.005
Steroid	13.5	0	11.5	0	0.022	10.5	0	8.4	0	0.026	12.4	0	13.4	0	0.011
Laboratory findings															
BUN (mg/dL)	15.1±5.7	0	15±5.2	0	0.051	14.1±5.3	0	13.7±4.9	0	0.082	14.7±5.5	2	14.4±5.1	2	0.073
Hemoglobin (g/dL)	13.1±1.9	0	13.3±1.8	0	0.063	12.8±1.9	0	12.7±1.9	0	0.037	12.9±1.8	0	12.9±1.9	0	0.023
Albumin (g/dL)	4.1±0.5	0	4.1±0.5	0	0.087	3.9±0.7	0	3.9±0.6	0	0.098	3.7±0.5	0	3.7±0.5	0	0.064
Cr (mg/dL)	0.9±0.3	0	0.9±0.2	0	0.047	0.8±0.2	0	0.8±0.2	0	0.066	0.8±0.2	0	0.8±0.2	0	0.018
Sodium (mmol/L)	140.4±2.7	0	140.5±2.6	0	0.123	140.2±3.0	0	140±3.0	0	0.118	140.3±2.6	0	140.3±2.6	0	0.127

2025_KSN_초록_Table1.png

Figure 1. The gradual augmentation effect of characteristic-matched synthetic data on real-world data of various sample sizes.

Three sets of synthetic data, generated from three data-rich hospitals, were used to augment the real-world data of six hospitals (three data-rich and three data-scarce). The left side of each graph's subtitle indicates the origin of the synthetic data, while the right side indicates the origin of the real-world data. Data-rich hospitals were augmented using synthetic data from the other two institutions, excluding their own. Synthetic data was selected through exact matching with real-world data. The y-axis represents the AUROC difference between the PO-AKI prediction model trained on augmented data and the model trained solely on the corresponding real-world patient data. **Abbreviation:** SNUH, Seoul National University Hospital; SNUBH, Seoul National University Bundang Hospital; AMC, Asan Medical Center; BMC, Seoul Metropolitan Government Boramae Medical Center, BMC; CAUH, Chung-Ang University Hospital; KNUH, Kangwon National University Hospital;