A Digital Twin Framework for ICU Delirium Using SAITS-Based Patient State Embeddings

Sung man Hong[1] and Sangjeong Ahn[2⋆]

[1] Department of Biomedical Informatics, Korea University College of Medicine, Seoul, Republic of Korea
skilove13@korea.ac.kr

[2] Department of Pathology, Korea University Anam Hospital, Seoul, Republic of Korea
vanitas80@korea.ac.kr (corresponding author)

Abstract. Intensive care unit (ICU) delirium affects 20–80% of critically ill patients and is associated with prolonged hospitalization, long-term cognitive impairment, and increased mortality; yet existing prediction models are largely single-cohort binary classifiers that cannot represent an evolving patient state or support simulation-based decision making. We present a **digital-twin-inspired framework for ICU delirium** that represents each patient as a continuous latent *state* embedding from a **Self-Attention Imputation for Time Series (SAITS)** model—trained jointly for imputation and delirium prediction on hourly electronic health record (EHR) data—and unifies prediction, risk-trajectory analysis, phenotyping, and similar-patient retrieval over that state. We benchmark SAITS on three large ICU cohorts—MIMIC-III (n=4,047), MIMIC-IV (n=25,697), and eICU-CRD (n=6,631; mean missingness ≈83%, and broadly multi-center)—against seven baselines spanning classical machine learning (LR, RF, LightGBM, XGBoost), deep tabular models (TabPFN v2, FT-Transformer), and an EHR transformer (TransEHR). SAITS is not the strongest discriminator overall—TabPFN and gradient-boosting trees lead on the well-observed MIMIC cohorts—but on the extreme-missingness eICU cohort it attains the **best F1 (0.651; +0.082 over LightGBM)** and best AUROC (0.765), indicating robustness when data are severely incomplete. Crucially, the 64-dimensional SAITS state embedding—which tree and in-context baselines do not provide—enables three digital-twin capabilities: **phenotype clustering** (four subgroups, delirium rates 6.3–57.1%, a 9-fold spread); **temporal risk trajectories** flagging high-risk patients within 1–2 hours; and **similar-patient retrieval** that recovers high-risk peers and surfaces a model false negative. The latent state further offers a natural substrate for future counterfactual intervention simulation—not yet implemented here—moving ICU delirium management beyond binary prediction toward clinical decision support.

Keywords: Digital Twin · ICU Delirium · Patient Embedding · Self-Attention Imputation · Electronic Health Records · Time-Series Missingness · Clinical Decision Support

⋆ Corresponding author.

1 Introduction

Delirium is an acute disturbance of brain function that occurs in 20–80% of intensive care unit (ICU) patients and is strongly linked to prolonged stay, extended ventilation, post-discharge cognitive impairment, and elevated mortality [1–3]. Early detection enables preventive intervention [5], yet clinical practice still relies on nurse-administered tools such as the Confusion Assessment Method for the ICU (CAM-ICU), which adds latency and workload [4].

EHR-based automated prediction is a promising way to close this gap, and many studies have applied machine and deep learning to it [6, 7], but existing work has three recurring limitations. **(i) Single-cohort scope:** models are typically trained and evaluated on one dataset, leaving external validity unaddressed. **(ii) Static binary framing:** predicting only delirium presence discards the temporal evolution of risk. **(iii) Under-treated missingness:** ICU time series are severely incomplete (≈83% in eICU-CRD), yet imputation is often a preprocessing afterthought.

From embedding to digital twin. A digital twin is a dynamic virtual replica that, synchronized with live data, supports state estimation, temporal evolution, scenario simulation, and decision support [22, 23]. We realise these on a single learned patient-state embedding: state estimation (risk), temporal evolution (risk trajectories), population positioning (phenotype clustering), and case-based reasoning (similar-patient retrieval), with the same state as a substrate for future counterfactual simulation. This is what separates such a state from a one-off embedding—it is queryable, comparable, and evolvable rather than a single prediction. Our framework realises the *representation* components of a twin; mechanism constraints and intervention simulation are left to future work, and twins specific to ICU delirium remain scarce.

Why SAITS. SAITS (Self-Attention Imputation for Time Series) [8] is an *existing* time-series imputation model that uses Diagonal-Masked Self-Attention (DMSA) to reconstruct missing values. Rather than proposing a new predictor, we *repurpose* it as a patient-state encoder: unlike tree ensembles or in-context tabular models (e.g. TabPFN), which emit only a scalar risk, SAITS produces a continuous representation of the multivariate time series. Its encoder output ($h^{(2)}$) places the patient's state in a 64-dimensional space, so the twin state comes *for free* from the same model that imputes and predicts.

Contributions.

1. **Digital-twin-inspired framework:** we represent each patient as a SAITS latent state and unify prediction, risk trajectories, phenotyping, and similar-patient retrieval over that state—a queryable, evolvable representation rather than a one-off classifier, offering a substrate for future intervention simulation.
2. **Multi-cohort benchmark:** eight models—classical ML, deep tabular, and EHR transformers—compared across three ICU cohorts (MIMIC-III/IV, eICU-CRD; 36,375 patients).

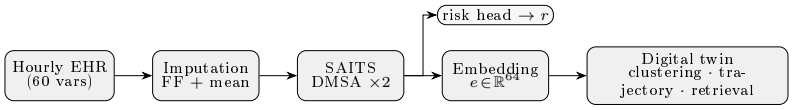

Fig. 1. SAITS-based digital twin framework. Imputed hourly EHR (forward-fill + global mean; mask M, gap Δ) is encoded by two DMSA stacks; $h^{(2)}$ feeds a delirium-prediction head and, after mean-pooling over observed steps, yields a 64-D patient embedding e that serves as the *digital twin state*—driving phenotype clustering, risk-trajectory monitoring, and similar-patient retrieval.

Table 1. ICU cohort characteristics (Setting A). Missingness is the mean fraction of unobserved variable–hour cells.

Cohort	Patients	Delirium+	Missingness	ICU type
MIMIC-III	4,047	24.2%	≈38%	Mixed ICU
MIMIC-IV	25,697	38.5%	≈40%	MICU/SICU/CCU
eICU-CRD	6,631	31.4%	≈83%	208 hospitals

3. **High-missingness robustness:** on eICU-CRD (83% missingness) SAITS attains the best F1 (and AUROC) among all models, indicating robustness of self-attention imputation to extreme incompleteness.
4. **Embedding-based error recovery:** in a case study, retrieval surfaces a high-risk false negative missed by the scalar classifier, suggesting a potential safety-net role that warrants systematic evaluation.

2 Method

2.1 Datasets and Preprocessing

We use three publicly available, credentialed ICU databases containing CAM-ICU records: **MIMIC-III** v1.4 [21] (2001–2012, single center), **MIMIC-IV** v2.2 [19] (2008–2019, multiple ICU types), and **eICU-CRD** v2.0 [20] (208 hospitals). Cohort sizes and characteristics are summarized in Table 1.

From hourly EHR records we extract up to 60 clinical variables spanning vital signs (8; e.g. HR, MAP, SpO_2, GCS), laboratory values (20; e.g. WBC, Na, BUN, glucose), medications (15; e.g. dexmedetomidine, haloperidol, lorazepam), severity scores (5: SOFA, APACHE II, LODS, SAPS, OASIS), and nursing assessments (12; e.g. RASS, pain, restraint). After removing variables with excessive missingness or duplication (RASS Goal, serum osmolality, SOFA without GCS), 53–60 variables remain per cohort.

Observation windows. Our primary experiment (**Setting A**) selects one CAM-ICU event per patient (the first positive assessment, or a random negative for never-positive patients) and uses the *8-hour window strictly preceding that assessment* as input to predict its outcome, giving one non-overlapping sample per

patient; no post-assessment measurements are used. Some window variables (e.g. sedation, nursing assessments) may partly reflect clinicians' emerging suspicion of delirium, so a leakage-controlled ablation excluding such variables is left to future work.

2.2 Missingness Handling

We compare two imputation strategies. **Median imputation** aggregates each window into mean/standard-deviation features, applies a median imputer, and standardizes (the standard prior-work baseline). **LOCF (forward-fill + global mean)** propagates the last observed value and fills leading gaps with the training global mean, preserving temporal trend. Both are applied *after* window aggregation, which is not equivalent to forward-filling the hourly series first. SAITS additionally performs **learned imputation**: this forward-fill initialization is refined by DMSA self-attention using explicit missingness masks and time gaps (Δ).

2.3 SAITS and Patient Embedding Extraction

Let $X \in \mathbb{R}^{T\times F}$ be a patient's multivariate series with observation mask $M \in \{0,1\}^{T\times F}$ and time-gap matrix Δ. SAITS stacks two DMSA blocks (Fig. 1). Diagonal masking blocks a time step from attending to itself via a diagonal mask M, forcing each step to be reconstructed from the others:

$$\mathrm{DMSA}(Q,K,V) = \mathrm{softmax}\Big(\tfrac{QK^\top}{\sqrt{d}} + M\Big)V, \quad M_{ii}{=}{-}\infty,\ M_{ij}{=}0\ (i{\neq}j). \tag{1}$$

The first stack ($h^{(1)}$) produces an initial imputation; the second stack ($h^{(2)}$) refines it and feeds the prediction head. Training combines the two standard SAITS reconstruction objectives—observed reconstruction ($\mathcal{L}_{\mathrm{ORT}}$) and masked imputation ($\mathcal{L}_{\mathrm{MIT}}$)—with a supervised delirium loss: $\mathcal{L} = \mathcal{L}_{\mathrm{ORT}} + \mathcal{L}_{\mathrm{MIT}} + \lambda\,\mathcal{L}_{\mathrm{cls}}$.

Patient embedding. We define the patient embedding by mean-pooling the second-stack representation over *observed* time steps only, minimizing the influence of missing positions:

$$e_i = \frac{1}{|\mathcal{O}_i|}\sum_{t\in\mathcal{O}_i} h_{i,t}^{(2)}, \qquad e_i \in \mathbb{R}^{64}, \tag{2}$$

where $\mathcal{O}_i$ is the set of observed steps for patient i. The SAITS configuration for Setting A is d_{model}=64, 4 heads, 2 layers, $T_{\max}$=8, F=53–60, totalling 153,645 parameters.

2.4 Digital Twin Capabilities

The embedding space $\{e_i\}$ supports three functions: **phenotype clustering** (KMeans k=4, visualized via t-SNE [15]/PCA); **temporal risk trajectories** (progressively longer prefixes yield an hour-by-hour risk curve for early warning); and **similar-patient retrieval** (cosine nearest neighbors $\mathrm{sim}(e_i,e_j) = e_i^\top e_j/(\|e_i\|\|e_j\|)$ returning comparable past patients and their outcomes).

3 Experiments

3.1 Evaluation Protocol

We compare eight models: four classical baselines—logistic regression (LR), random forest (RF) [14], LightGBM [12], XGBoost [13]—two deep tabular models (TabPFN v2 [9]; FT-Transformer [10]), the self-supervised EHR transformer TransEHR [11], and SAITS [8]. Classical models use 5-fold stratified cross-validation. We report AUROC (discrimination), F1 at the optimal threshold, and sensitivity (the delirium detection rate, the key screening metric); cluster quality is assessed by the silhouette [16], Davies-Bouldin [17], and Calinski-Harabasz [18] indices, and retrieval quality by Recall@k and mean average precision (mAP). The comparison is not fully input-matched (tabular/tree baselines use window-level summary features, whereas SAITS and TransEHR receive the full sequence), so it reflects practical pipelines rather than architectures under identical inputs.

3.2 Delirium Prediction across Cohorts

Table 2 compares test-set AUROC across the eight models. On the two lower-missingness MIMIC cohorts, deep-tabular TabPFN v2 and the gradient-boosting trees lead (AUROC up to 0.894), and SAITS is mid-pack; we do not claim state-of-the-art discrimination. The picture changes under the ≈83% missingness of eICU-CRD: **SAITS attains the best F1 there (**0.651**; +**0.082 **over LightGBM)** and the best AUROC (0.765, a narrower +0.027 margin), indicating that its self-attention imputation is the most robust when data are severely incomplete. Notably, despite the highest average AUROC, TabPFN has the lowest sensitivity of the models (as low as 0.386 on eICU), unsuitable for delirium screening where a missed case delays care. We adopt SAITS not for peak discrimination but because it alone yields the continuous patient-state embedding underlying our digital twin (Sec. 3.4).

Table 2. Model comparison: test-set AUROC (Setting A) across the three cohorts. Best per column **bold**. On the extreme-missingness eICU cohort, SAITS attains the best AUROC.

Type	**Model**	M-III	M-IV	eICU	**Avg.**
Classical	LR	0.809	0.823	0.689	0.773
	RF [14]	0.868	0.873	0.735	0.826
	LightGBM [12]	0.879	0.893	0.738	0.837
	XGBoost [13]	0.876	0.892	0.733	0.834
Deep tab.	TabPFN v2 [9]	**0.884**	**0.894**	0.748	**0.842**
	FT-Transformer [10]	0.808	0.872	0.737	0.806
EHR tr.	TransEHR [11]	0.822	0.862	0.702	0.795
Ours	**SAITS** [8]	0.835	0.881	**0.765**	0.827

3.3 Effect of Missingness Handling

Table 3 contrasts median and LOCF imputation on MIMIC-IV (LightGBM). Median gives marginally higher AUROC (+0.010), but LOCF improves sensitivity (+0.036) and F1 (+0.008) by preserving temporal trend, trading some specificity (−0.061). Since missed delirium is costlier than a false positive in ICU monitoring, this trade favours LOCF for screening *in this cohort*. We do not generalise beyond a single cohort, model, and window length.

Table 3. Median vs. LOCF (forward-fill + global mean) imputation on MIMIC-IV Setting A (LightGBM). Δ is LOCF − Median.

Metric	Median	LOCF	Δ
AUROC	**0.8930**	0.8832	−0.010
AUPRC	0.8301	**0.8395**	+0.009
F1	0.7774	**0.7856**	+0.008
Sensitivity	0.8279	**0.8640**	+0.036
Specificity	**0.8107**	0.7500	−0.061

3.4 Digital Twin Analysis (MIMIC-III, Setting A)

We extract embeddings for all 4,047 MIMIC-III patients (input 4047×8×53, observation rate 60.3%), yielding a 4047×64 representation. Delirium and non-delirium patients separate structurally, with predicted risk varying smoothly across the t-SNE manifold (Fig. 2, bottom-right), indicating clinically meaningful rather than arbitrary structure.

Phenotype clustering. KMeans (k=4) yields four phenotypes with delirium rates spanning 6.3%–57.1% (a 9.1× difference; Table 4; Fig. 2, bottom-centre/right); only high-risk Cluster 2 exceeds 0.5. Validity is modest but non-trivial (silhouette 0.157, Davies-Bouldin 2.17, Calinski-Harabasz 746.3)—soft boundaries consistent with a delirium spectrum. Delirium patients lie farther from their centroid (Fig. 2, top-right), i.e. are more heterogeneous, and risk-correlated signal is spread across many embedding dimensions ($|\rho| = 0.6$–0.8; Fig. 2, bottom-left), suggesting composite clinical patterns.

Temporal risk trajectory. Feeding progressively longer prefixes of each patient's series yields an hour-by-hour risk curve (Fig. 2, top-centre). Delirium patients cross the 0.5 threshold within the first 1–2 observation hours, whereas non-delirium patients remain below 0.3 across the full 8-hour window. This early separation suggests potential for a pre-emptive alert ahead of a nurse-administered CAM-ICU assessment, subject to prospective validation.

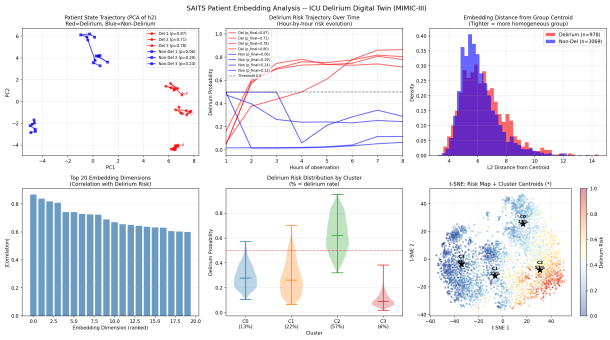

Fig. 2. SAITS patient-embedding analysis (MIMIC-III, Setting A). **Top:** PCA state trajectories; hour-by-hour risk (delirium crosses 0.5 at 1–2 h); L2 distance to centroid (delirium more heterogeneous). **Bottom:** top-20 risk-correlated dimensions; risk per KMeans cluster (%= delirium rate); t-SNE risk map with centroids ($\star$).

Table 4. Delirium phenotype clusters (KMeans, k=4; silhouette 0.157, Davies-Bouldin 2.17, Calinski-Harabasz 746.3). Cluster 2 vs. Cluster 3: 57.1% vs. 6.3% delirium (9.1×).

Cluster	Patients	Delirium rate	Mean risk	Phenotype
0	931	13.0%	0.290	Low–moderate risk
1	906	22.0%	0.283	Moderate / borderline
2	1021	57.1%	0.629	High risk (intervention priority)
3	1189	6.3%	0.101	Low risk / safe

Similar-patient retrieval. We rank neighbors by cosine similarity in the normalised state space. Cohort-level Top-k retrieval attains Recall@5 0.754, Recall@10 0.753, and mAP 0.637, but is asymmetric: accurate for non-delirium queries (Recall@5 0.851, mAP 0.757) yet much harder for the minority delirium class (0.447, 0.257). This lower performance likely reflects the greater clinical heterogeneity of delirium patients in the embedding space. As a case study (Fig. 3), patient 234196—predicted non-delirium by the scalar classifier—retrieves five delirium neighbors with high embedding-space risk (0.768), illustrating how retrieval could surface a false negative the point prediction misses; establishing a reliable safety net requires systematic evaluation, left to future work.

4 Discussion

Discrimination vs. robustness. These results reframe SAITS's role: it is not a universally stronger discriminator—TabPFN and boosting lead on the well-observed MIMIC cohorts—but under extreme missingness its imputation is the most robust (best eICU F1, and AUROC by a small margin), and, unlike the baselines, it yields the patient-state embedding that powers the downstream analyses. The gap on the smaller MIMIC cohorts is consistent with attention-based sequence

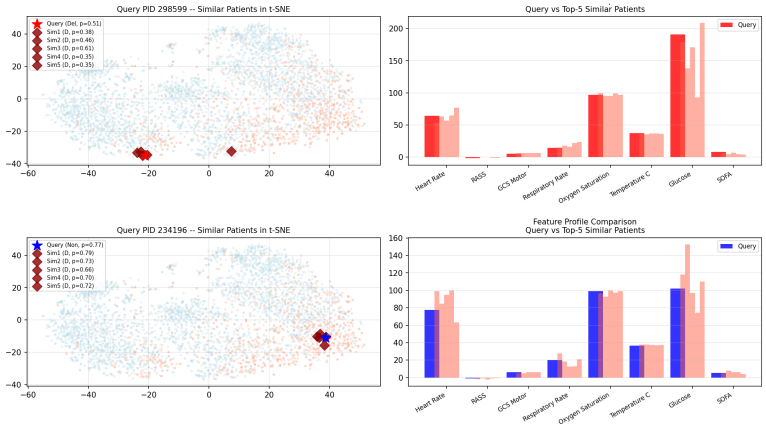

Fig. 3. Similar-patient retrieval (cosine Top-5). **Top:** query 298599 (delirium) retrieves five delirium neighbors. **Bottom:** query 234196, predicted non-delirium, retrieves five delirium neighbors (p=0.66–0.79)—a false negative surfaced by the embedding. Right panels compare query (solid) vs. neighbor feature profiles.

models being more data-hungry than gradient-boosted trees, whose advantage is largest when features are well observed and samples are limited. As a persistent, comparable patient state, the embedding also supports what a scalar risk cannot: longitudinal monitoring, phenotype-tailored care planning, and cross-hospital state comparison in a shared latent space, and it offers a substrate for counterfactual intervention simulation—the direction that most fully realises the digital-twin vision.

Limitations. First, SAITS does not maximise AUROC; its advantage is confined to robustness under heavy missingness and to the embedding, and the eICU margin is small. Second, retrieval is correlational, weaker for the minority delirium class (Recall@5 0.45), and counterfactual simulation is unvalidated. Third, embedding analysis is limited to MIMIC-III with soft clusters (silhouette 0.157). Finally, we omit confidence intervals, calibration, and decision-curve analysis—all targets for the extended study.

5 Conclusion

We presented a digital-twin-inspired framework that casts each ICU patient as a SAITS latent state unifying prediction, risk trajectories, phenotyping, and retrieval. Across three cohorts (36,375 patients), SAITS is the most robust under extreme missingness (best eICU F1 and AUROC) though not the overall best discriminator. Future work will add the twin's simulation arm and broader validation, positioning the SAITS latent state as a practical foundation for clinical digital twins.

Ethical Approval and Informed Consent. This study used the MIMIC-III, MIMIC-IV, and eICU Collaborative Research Database, which are publicly available, fully de-identified critical care databases distributed through PhysioNet. Data collection for MIMIC-III and MIMIC-IV was approved by the Institutional Review Boards of the Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, and the eICU-CRD was granted exemption by the Institutional Review Board of the Massachusetts Institute of Technology. As all records are de-identified in accordance with HIPAA Safe Harbor provisions, the requirement for individual informed consent was waived, and no additional institutional review board approval was required for this secondary analysis. The authors completed the required training and signed the data use agreements prior to accessing the data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ely, E.W., et al.: Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. JAMA **291**(14), 1753–1762 (2004)
2. Pandharipande, P.P., et al.: Long-term cognitive impairment after critical illness. N. Engl. J. Med. **369**(14), 1306–1316 (2013)
3. Girard, T.D., et al.: Delirium as a predictor of long-term cognitive impairment in survivors of critical illness. Crit. Care Med. **38**(7), 1513–1520 (2010)
4. Ely, E.W., et al.: Evaluation of delirium in critically ill patients: validation of the CAM-ICU. Crit. Care Med. **29**(7), 1370–1379 (2001)
5. Barr, J., et al.: Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the ICU. Crit. Care Med. **41**(1), 263–306 (2013)
6. Adamson, J., et al.: Machine learning for delirium prediction in the ICU. J. Crit. Care **65**, 1–8 (2021)
7. Wong, A., et al.: External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern. Med. **181**(8), 1065–1070 (2021)
8. Du, W., Cote, D., Liu, Y.: SAITS: Self-attention-based imputation for time series. Expert Syst. Appl. **219**, 119619 (2023)
9. Hollmann, N., Müller, S., Eggensperger, K., Hutter, F.: TabPFN: a transformer that solves small tabular classification problems in a second. In: ICLR (2023)
10. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. In: NeurIPS **34**, 18932–18943 (2021)
11. Xu, Y., Xu, S., Ramprassad, M., Tumanov, A., Zhang, C.: TransEHR: self-supervised transformer for clinical time series data. In: Proc. Machine Learning for Health (ML4H), PMLR **225**, pp. 623–635 (2023)
12. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: a highly efficient gradient boosting decision tree. In: NeurIPS **30**, pp. 3146–3154 (2017)
13. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proc. 22nd ACM SIGKDD, pp. 785–794 (2016)

14. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
15. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
17. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-1**(2), 224–227 (1979)
18. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**(1), 1–27 (1974)
19. Johnson, A.E.W., et al.: MIMIC-IV, a freely accessible electronic health record dataset. Sci. Data **10**, 1 (2023)
20. Pollard, T.J., et al.: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci. Data **5**, 180178 (2018)
21. Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)
22. Corral-Acero, J., et al.: The 'digital twin' to enable the vision of precision cardiology. Eur. Heart J. **41**(48), 4556–4564 (2020)
23. Bruynseels, K., Santoni de Sio, F., van den Hoven, J.: Digital twins in health care: ethical implications of an emerging engineering paradigm. Front. Genet. **9**, 31 (2018)