CalTwin: Calibrated Shift-Robust Medical World Models via Fisher-Information Regularization

Behraj Khan[1,2][0000−0003−0985−9543], Tahir Qasim Syed[2][0000−0003−0638−9689], Shabir Ahmad[3][0000−0002−8788−2717], Vincent Martin Vigneron[4][0000−0001−5917−6041], and Syed Ahmad Chan Bukhari[1][0000−0002−6517−5261]

[1] St. John's University, New York, USA
[2] Institute of Business Administration Karachi, Pakistan
[3] Gachon University, Seongnam, South Korea
[4] Université Paris-Saclay, France

Abstract. Medical world models aim to learn latent representations of patient physiology and transition dynamics that predict future states under temporal evolution and clinical interventions. However, their reliability is challenged by two major issues: distribution shift caused by fragmented multi-site clinical data, and confidence misalignment during multi-step forecasting. We present **CalTwin**, a unified regularisation framework that combines Fisher-Information-based parameter regularisation for mitigating fragment-level shift with a Confidence Misalignment Penalty for improving prediction reliability. The proposed formulation adapts these principles to a latent transition predictor of a medical world model, where Fisher regularisation is applied to the transition likelihood rather than classification objectives, and confidence calibration is performed through a shared clinical prediction head. CalTwin is evaluated on multi-hospital ICU time-series data using the PhysioNet 2019 Sepsis Challenge, with sequential hospital fragments for training and an unseen hospital for out-of-distribution evaluation. CalTwin reduces OOD next-step latent-state mean squared error by 9.1% compared with an unregularised baseline, with Fisher regularisation providing the primary contribution. Confidence calibration improvements are modest, with expected calibration error reduced by 0.7% using CalTwin. An additional evaluation on eICU-CRD demonstrates that the contribution of each component depends on the characteristics of the clinical distribution shift. These results highlight the potential of combining shift-robust learning and confidence-aware modelling for medical world models while motivating further validation on larger-scale multimodal clinical datasets, including medical imaging.

1 Introduction

Medical world models learn compact representations of patient or organ states and transition functions modelling their evolution under temporal progression,

interventions, or new observations, typically through $p(s_{t+1} \mid s_t, a_t)$. Capabilities range from temporal prediction (L1) and action-conditioned simulation (L2) to counterfactual reasoning (L3) and closed-loop planning (L4) [8], though reliable L3–L4 systems remain challenging. Medical digital twins offer a potential framework by integrating patient data, in-silico models, clinical interfaces, and synchronization mechanisms, but robust multimodal integration and trustworthy learning remain unresolved [11, 10].

Two factors critically limit reliability. First, clinical data are fragmented across institutions, scanners, protocols, and time, with privacy constraints motivating federated or sequential learning; each fragment may therefore follow a distinct distribution $P_k(s_t)$, producing fragment-level covariate shift that can impair cross-hospital generalization [16], with such errors compounding during multi-step rollouts. Conventional importance weighting, $P_{\text{dep}}(s)/P_{\text{train}}(s)$ [12], is impractical when deployment distributions are unknown and training distributions evolve. FIM-based regularization instead preserves information from previous fragments while adapting to new distributions; we adapt the FIcsR framework [5, 3] to latent transition dynamics.

Second, autoregressive forecasting creates confidence misalignment: training uses teacher-forced ground-truth states, whereas deployment conditions on model-generated states, so prediction errors propagate across the horizon while confidence is estimated along increasingly uncertain trajectories. Since neural networks can be systematically overconfident [2] and post-hoc calibration may degrade under distribution shift [13], reliability requires jointly accounting for confidence and distributional variation.

We therefore study a unified Fisher-Information-based framework for robust, calibrated medical world-model learning. FIcsR [5, 3] is adapted for cross-fragment robustness, while CalShift [4] and its confidence modification penalty (CMP) improve predictive calibration. We deliberately frame this as a controlled feasibility study a minimal encoder transition predictor evaluated under one-step, teacher-forced prediction, rather than a full L2–L4 world model and use multi-hospital ICU time-series data as a proxy for fragmented clinical learning to assess whether these mechanisms provide a principled foundation for reliable latent dynamics modeling.

2 Related Work

Medical digital twins and world models. Medical digital twins extend digital-twin concepts to patient-specific modelling, but clinical translation remains constrained by multimodal integration, privacy, and trustworthy learning [11, 10]. Medical world models similarly learn latent physiological states and transition dynamics for prediction, simulation, and planning across imaging, EHR, and surgical domains [8]. We focus on the latent transition predictor, where distributional robustness and calibrated uncertainty are critical.

Covariate shift correction. Classical methods correct covariate shift via train–test density-ratio reweighting [5], impractical when fragment and deployment

distributions both evolve. FIcsR and its C^3 variant instead use FIM-based regularisation to approximate distributional divergence in parameter space, enabling incremental adaptation without retaining previous data [5, 3]; we extend this to medical transition dynamics.

Confidence calibration under shift. Distribution shift and limited-data adaptation can induce overconfident predictions [4]. CalShift combines FIM regularisation with a Confidence Misalignment Penalty (CMP) to suppress such errors [4], but its application to latent dynamics in medical world models remains unexplored.

Distribution-free calibration. Conformal methods offer post-hoc, distribution-free calibration via reweighted nonconformity scores [15], calibrated multi-horizon regions [14], and drift-aware coverage tracking [1]. Unlike CMP, these guarantee coverage for a fixed model but do not address parameter drift during fragment-wise training; we view the two as complementary and leave conformal baselines to future work.

Clinical multi-hospital shift. Federated clinical prediction studies show cross-hospital demographic and practice heterogeneity substantially affects performance [16], motivating our use of multi-hospital ICU data as a proxy for fragmented clinical learning and cross-site shift.

3 Proposed Method

Problem formulation. Following the medical world-model formulation [8], an encoder $\mathcal{E}$ maps clinical observations x_t to latent states $s_t = \mathcal{E}(x_t)$, and a transition model predicts $p_\theta(s_{t+1}|s_t, a_t)$. Training data consist of K fragments $\{B_1, \ldots, B_K\}$, where each fragment may correspond to a hospital, protocol, or time period with distribution $\hat{P}_k(s_t)$. Our objective is to improve cross-fragment generalisation while maintaining reliable confidence during autoregressive prediction.

Fisher-information regularisation. To reduce fragment-induced distribution shift, we adapt FIcsR [5, 3] by regularising transition-model parameters using accumulated Fisher Information. The KL approximation is $D_{\mathrm{KL}}(P(\theta|B_k)||Q(\hat{\theta})) \approx \frac{1}{2}(\hat{\theta}-\theta)^T I(\theta)(\hat{\theta}-\theta)$, where the empirical Fisher matrix for the transition predictor is $I_e(\theta) = \frac{1}{N}\sum_i g_i g_i^T, \quad g_i = \frac{\partial \log p_\theta(s_{i,t+1}|s_{i,t},a_{i,t})}{\partial \theta}$. Unlike previous classification settings, the gradients are computed from the transition likelihood rather than categorical loss. A running Fisher estimate is updated as

$$\hat{I}^{(k)} = \alpha \hat{I}^{(k-1)} + (1-\alpha) I_e^{(k)}. \tag{1}$$

The shift-robust objective becomes

$$\mathcal{L}_{shift} = -\log p_\theta(s_{t+1}|s_t, a_t) + \lambda_1 (\hat{\theta}-\theta)^T \hat{I}^{(k-1)} (\hat{\theta}-\theta). \tag{2}$$

$I_e^{(k)}$ is diagonal, estimated from all mini-batches in fragment B_k at the end of its final epoch, using one Monte Carlo sample per transition drawn from $p_\theta(s_{t+1} \mid s_t, a_t)$.

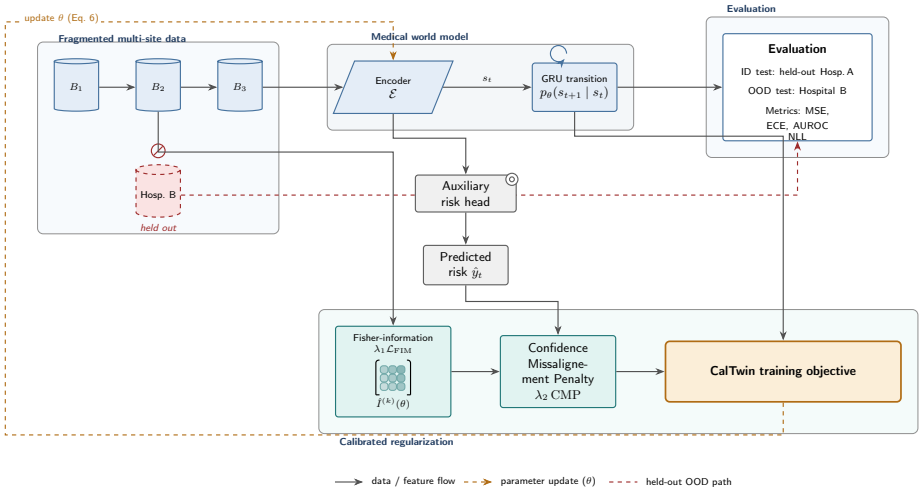

Fig. 1. CalTwin training pipeline. Sequential hospital fragments B_1, B_2, B_3 train the shared encoder $\mathcal{E}$ and GRU transition predictor, while a held-out hospital is reserved for OOD evaluation. CMP regularises an auxiliary risk head, while diagonal Fisher information anchors encoder/transition parameters to prior fragments. Both penalties and transition likelihood form the CalTwin objective (Eq. 4), which updates θ.

This is the same second-order KL approximation used by Elastic Weight Consolidation [6], with the classification log-likelihood replaced by the transition log-likelihood; we make this relationship explicit rather than treating Eq. 4 as a novel penalty in its own right, and we discuss the consequence for our claimed contribution in Sec. 5.

Confidence Misalignment Penalty. Autoregressive forecasting creates a mismatch between teacher-forced training and self-conditioned inference. To penalize confidence degradation, we adapt the Confidence Misalignment Penalty (CMP) from [4]. For a discrete auxiliary prediction head $q_\theta(y \mid s_t)$ sharing the latent representation, define $\mathcal{Y}^{>}(s_t, y) = \{y' \neq y : q_\theta(y' \mid s_t) > q_\theta(y \mid s_t)\}$. The CMP is

$$\mathrm{CMP}(s_t, y) = \begin{cases} 0, & \mathcal{Y}^{>}(s_t, y) = \varnothing, \\ \dfrac{q_\theta(y \mid s_t)}{\sum_{y' \in \mathcal{Y}^{>}(s_t, y)} q_\theta(y' \mid s_t)}, & \text{otherwise.} \end{cases} \tag{3}$$

The first case is required because the ratio is undefined when the true class already ranks first. We note that the ratio form requires verification against the original CalShift formulation [4]: as written, increasing competing probability mass decreases CMP, which is inconsistent with minimizing $+\lambda_2$CMP as a misalignment penalty. We resolve this in Sec. 5: as defined, CMP decreases with competing probability mass and is therefore a misalignment *reward*; we correct

it to $\mathrm{CMP}'(s_t, y) = \sum_{y' \in \mathcal{Y}^>(s_t,y)} q_\theta(y' \mid s_t)/q_\theta(y \mid s_t)$, which increases monotonically with competitor mass and is consistent with minimizing $\lambda_2 \mathrm{CMP}'$.
Because the transition model is continuous, CMP is applied to a shared auxiliary clinical prediction head rather than directly to $p_\theta(s_{t+1} \mid s_t, a_t)$.
Unified Objective. The final objective jointly optimizes transition learning, shift robustness, and confidence calibration:

$$\mathcal{L}_{\mathrm{CalTwin}} = -\log p_\theta(s_{t+1} \mid s_t, a_t) + \lambda_1(\hat{\theta}-\theta)^T \hat{I}^{(k-1)}(\hat{\theta}-\theta) + \lambda_2 \mathrm{CMP}(s_t, y_t). \quad (4)$$

This objective extends FIM-based shift correction and confidence calibration to medical world-model dynamics. The two penalties are evaluated jointly as an empirical hypothesis rather than assumed to be complementary: Table 1 shows that the combined objective improves transition accuracy over either penalty alone but is inferior to CMP-only on both OOD ECE and OOD AUROC. Thus, "unified" denotes joint optimization, not necessarily joint benefit.

4 Experiments

We evaluate CalTwin as a feasibility study for robust medical world models under fragmented clinical data and cross-site distribution shift, using multi-hospital ICU time-series data as a proxy setting; imaging-based validation is outside scope. Evaluation is one-step-ahead under teacher forcing; closed-loop autoregressive evaluation remains an open experiment.
Dataset and fragmentation. We use the PhysioNet 2019 Sepsis Challenge dataset [9], containing hourly ICU measurements and binary sepsis labels. Hospital A is divided chronologically into three fragments ($K = 3$, 1,200 subjects each), with the final 15% of each reserved for ID testing; Hospital B is held out for OOD evaluation. Sepsis prevalence across the three Hospital-A fragments is 8.8%, 9.6%, and 9.5%. Fragment separability was not independently verified via a domain classifier or MMD, so these fragments should not yet be read as a confirmed covariate-shift benchmark.
Each trajectory has seven vital-sign channels (heart rate, SpO_2, temperature, systolic/mean/diastolic blood pressure, respiratory rate), padded or truncated to 24 hourly steps and standardized per patient, with missing values forward/backward filled and fully missing channels replaced by reference values.
Model and training. The transition model encodes the seven-dimensional input into a 32-dimensional latent state, followed by a GRU with a diagonal Gaussian output $p_\theta(s_{t+1} \mid s_t)$. A shared auxiliary MLP predicts sepsis onset from the latent representation but is optimized only through the CMP term, without a standard supervised loss, limiting the interpretability of its calibration and AUROC results. The model has 24,385 trainable parameters.
Models are trained sequentially over the three Hospital-A fragments using Adam ($lr = 10^{-3}$, weight decay 10^{-4}), batch size 32, gradient clipping, and 15 epochs per fragment. The empirical Fisher matrix is updated via exponential moving average ($\alpha = 0.9$). Hyperparameters $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$ were selected by

Table 1. Performance on PhysioNet 2019 Sepsis Challenge. Lower MSE/ECE and higher AUROC are better. A constant base-rate predictor is not included as an ECE baseline. Values are mean ± std over 5 seeds; bold indicates the best mean.

Method	MSE-ID	MSE-OOD	ECE-ID	ECE-OOD	AUROC-ID	AUROC-OOD
Baseline	0.24 ± 0.02	0.26 ± 0.07	0.42 ± 0.03	0.44 ± 0.04	0.50 ± 0.00	0.50 ± 0.03
FIM-only	0.22 ± 0.05	0.24 ± 0.03	0.42 ± 0.08	0.44 ± 0.04	0.50 ± 0.00	0.50 ± 0.01
CMP-only	0.24 ± 0.01	0.26 ± 0.06	0.42 ± 0.01	$\mathbf{0.44} \pm 0.04$	0.57 ± 0.03	$\mathbf{0.50} \pm 0.01$
CalTwin	$\mathbf{0.22} \pm 0.02$	$\mathbf{0.24} \pm 0.05$	0.42 ± 0.02	0.44 ± 0.03	$\mathbf{0.57} \pm 0.03$	0.48 ± 0.02

Table 2. Performance on eICU-CRD Demo using the same evaluation protocol. Values are mean ± std over 5 seeds; bold indicates the best mean.

Method	MSE-ID	MSE-OOD	ECE-ID	ECE-OOD	AUROC-ID	AUROC-OOD
Baseline	0.15 ± 0.04	0.17 ± 0.07	0.49 ± 0.02	0.48 ± 0.03	0.50 ± 0.07	0.50 ± 0.01
FIM-only	0.15 ± 0.09	0.17 ± 0.04	0.49 ± 0.02	0.48 ± 0.03	0.50 ± 0.00	0.50 ± 0.09
CMP-only	0.14 ± 0.07	$\mathbf{0.16} \pm 0.08$	$\mathbf{0.38} \pm 0.09$	$\mathbf{0.37} \pm 0.08$	0.26 ± 0.04	$\mathbf{0.84} \pm 0.03$
CalTwin	$\mathbf{0.14} \pm 0.05$	0.16 ± 0.03	0.43 ± 0.02	0.43 ± 0.05	0.34 ± 0.04	0.84 ± 0.01

grid search over $\{0.01, 0.1, 1, 10\}$ on the ID validation split of fragment 3, using OOD MSE and OOD ECE respectively as selection criteria. We compare four regimes: Baseline, FIM-only, CMP-only, and CalTwin (joint).

We report next-step latent MSE, ECE, and auxiliary-head AUROC on ID/OOD splits, together with Brier score and NLL as complementary proper scoring rules, averaged over 5 seeds (mean ± std). Latent MSE is not invariant to the representation learned by each regularized encoder, so results indicate latent transition improvement rather than confirmed observation-space forecasting gains.

Reproducibility. Fragments are ordered chronologically within each hospital, with a fixed 1:1 fragment-to-hospital mapping in both PhysioNet and eICU-CRD. All hyperparameters (λ_1, λ_2, α, learning rate, epochs per fragment) are fixed across datasets, not re-tuned per dataset.

Results. Figure 2 complements Table 1 with trajectory-level errors, calibration, training dynamics, and OOD reliability. FIM-based methods generally reduce latent transition error, with CalTwin achieving the lowest ID and OOD MSE; calibration gains are limited, as CMP-only achieves lower OOD ECE and higher OOD AUROC than CalTwin. The reliability diagram suggests the auxiliary head produces near-constant probabilities well above the OOD sepsis frequency, consistent with the absence of a supervised loss.

CalTwin reduces OOD latent MSE by 9.1% relative to baseline, with FIM-only accounting for most of this reduction; adding CMP yields a further 2.3-point reduction in OOD MSE. Conversely, adding FIM to CMP degrades both OOD ECE and AUROC. Thus FIM primarily benefits latent transition prediction on PhysioNet, while CMP provides the stronger calibration signal. OOD AUROC stays near chance for all methods, so confidence transfer under cross-hospital shift is not established.

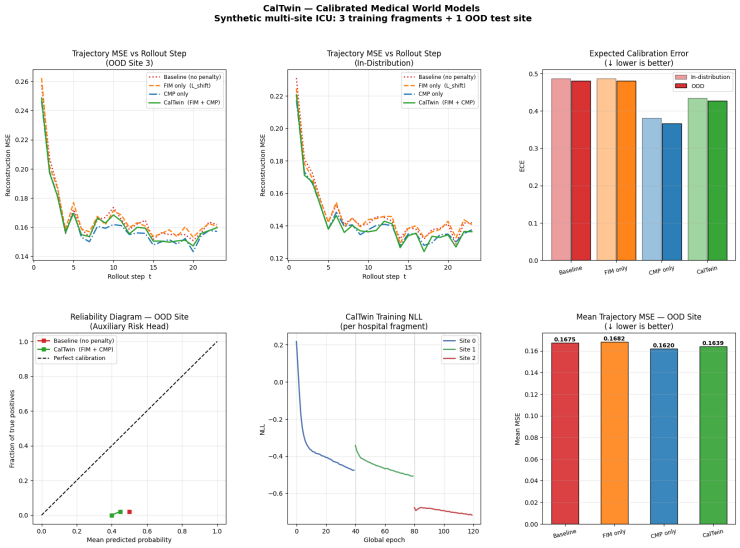

Fig. 2. PhysioNet 2019 results (single seed). Top: next-step latent MSE and ECE; bottom: OOD reliability, sequential training loss, and mean OOD MSE.

Additional validation on eICU-CRD Demo. We evaluate CalTwin on the eICU Collaborative Research Database Demo [7] under the same protocol, using three hospitals for sequential training and one held out for OOD evaluation.
On eICU, CMP-only is strongest on four of six metrics OOD MSE, both ECE measures, and OOD AUROC. CalTwin achieves the best ID MSE but again trails CMP-only on calibration and OOD discrimination; FIM-only gives negligible or negative transition gains over baseline. Across both datasets, CMP-only consistently dominates CalTwin on calibration, while FIM's effect on transition prediction is dataset-dependent. These results support joint optimization as feasible but do not establish that combining FIM and CMP is uniformly beneficial.

5 Discussion

The results clarify CalTwin's potential and limitations. First, the combined objective is not consistently superior for calibration: across both datasets, CMP-only achieves better or equal OOD ECE and AUROC than CalTwin (Tables 1–2), so the hypothesized benefit of combining FIM-based shift regularization with CMP is unsupported for calibration. A possible explanation is that FIM constrains the shared representation $s_t = \mathcal{E}(x_t)$, limiting the auxiliary head's ability to adapt its confidence estimates; this remains an untested hypothesis.
FIM's effect on transition prediction is dataset-dependent: FIM-only reduces PhysioNet OOD latent MSE but gives negligible or slightly negative change on eICU. MSE is also computed in an encoder-dependent latent space, not validated

in observation space, so these differences should be read as within-architecture comparisons rather than definitive forecasting-fidelity gains. We do not evaluate whether latent-MSE improvements translate to clinically actionable gains (e.g., earlier sepsis-onset detection); this requires observation-space decoding and outcome-linked metrics, left to future work. The auxiliary head also fails to establish reliable OOD calibration. Although CMP-trained models reach AUROC ≈ 0.57 on PhysioNet ID data, the OOD reliability diagram is consistent with near-constant predictions. Since the head is trained without a standard supervised loss, its weak, inconsistent AUROC gives limited evidence for confidence transfer under shift, and the absence of a constant base-rate predictor prevents meaningful interpretation of absolute ECE values.

The assumed fragment-level distribution shift is likewise unestablished: sepsis prevalence varies by less than one percentage point across the PhysioNet-A fragments, and no domain-classifier or MMD analysis was performed, so the observed FIM benefit may reflect generic parameter regularization rather than correction of verified covariate shift. The FIM term is also closely related to Elastic Weight Consolidation (EWC) [6] Eq. 4 uses the same Fisher-weighted quadratic anchoring principle, applied here to transition likelihood rather than classification loss and since EWC was not included as a baseline, we cannot separate our formulation's contribution from standard Fisher-based regularization.

Several limitations remain. Evaluation is one-step-ahead under teacher forcing and does not test the closed-loop error accumulation motivating the autoregressive calibration argument the most important missing experiment. Horizon-specific calibration is consequently not reported. CMP is applied through a binary auxiliary head since a principled continuous-output formulation for latent dynamics remains unresolved. Experiments use tabular ICU vital signs rather than imaging or multimodal data, and omit EWC, conformal prediction, and temperature-scaling baselines, restricting the strength and generality of the current empirical conclusions.

6 Conclusion

We introduced CalTwin, combining Fisher-information-based shift regularization with a Confidence Misalignment Penalty for latent transition prediction in medical world models. Across PhysioNet 2019 and eICU-CRD, Fisher regularization improves latent-space transition prediction to a dataset-dependent extent, but the combined objective is consistently outperformed by CMP-only on OOD calibration and discrimination the central empirical finding, indicating the two components do not exhibit the expected complementary benefit under the current setup. A plausible explanation is that FIM anchoring constrains the shared representation and thereby limits CMP-based confidence adaptation, though this mechanism remains unverified. Verified distribution-shift diagnostics, observation-space evaluation, and closed-loop autoregressive rollouts remain necessary before extending the approach to larger multimodal medical world models.

References

1. Gibbs, I., Candes, E.: Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems **34**, 1660–1672 (2021)
2. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
3. Khan, B., Mirza, B., Syed, T.: Causal covariate shift correction using fisher information penalty. arXiv preprint arXiv:2502.15756 (2025)
4. Khan, B., Qureshi, R., Durrani, N.M., Syed, T.Q.: Confidence-calibrated covariate shift correction for few-shot classification in vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6511–6523 (2025)
5. Khan, B., Syed, T.Q., Mirza, B., Durrani, N.M.: Mitigating covariate shift in non-colocated data with learned parameter priors. Neurocomputing p. 131656 (2025)
6. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
7. Pollard, T.J., Johnson, A.E., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eicu collaborative research database, a freely available multi-center database for critical care research. Scientific data **5**(1), 180178 (2018)
8. Qazi, M.A., Nadeem, M., Yaqub, M.: Beyond generative ai: World models for clinical prediction, counterfactuals, and planning. arXiv preprint arXiv:2511.16333 (2025)
9. Reyna, M.A., Josef, C., Seyedi, S., Jeter, R., Shashikumar, S.P., Westover, M.B., Sharma, A., Nemati, S., Clifford, G.D.: Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In: 2019 Computing in Cardiology (CinC). pp. Page–1. IEEE (2019)
10. Rudsari, H.K., Tseng, B., Zhu, H., Song, L., Gu, C., Roy, A., Irajizad, E., Butner, J., Long, J., Do, K.A.: Digital twins in healthcare: a comprehensive review and future directions. Frontiers in Digital Health **7**, 1633539 (2025)
11. Sadée, C., Testa, S., Barba, T., Hartmann, K., Schuessler, M., Thieme, A., Church, G.M., Okoye, I., Hernandez-Boussard, T., Hood, L., et al.: Medical digital twins: enabling precision medicine and medical artificial intelligence. The Lancet Digital Health **7**(7) (2025)
12. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference **90**(2), 227–244 (2000)
13. Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J.V., Ren, J., Nado, Z.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift (2019)
14. Stankeviciute, K., M Alaa, A., Van der Schaar, M.: Conformal time-series forecasting. Advances in neural information processing systems **34**, 6216–6228 (2021)
15. Tibshirani, R.J., Foygel Barber, R., Candes, E., Ramdas, A.: Conformal prediction under covariate shift. Advances in neural information processing systems **32** (2019)
16. Zhu, H., Bai, J., Li, N., Li, X., Liu, D., Buckeridge, D.L., Li, Y.: Fedweight: mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. npj Digital Medicine **8**(1), 286 (2025)