

Multimodal Generative AI for Missing Physiological Data Imputation in Long-Term Health Monitoring

Rohit Basu

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
rbasu@missouri.edu

Emile C. Cox

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
emilecox654@buffalo.edu

Larry Ortega

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence,
KS, USA.
larryortega83@ku.edu

Abstract

Long-term health monitoring systems increasingly rely on continuous streams of multimodal physiological signals produced by wearable and ambient sensors. These signals include electrocardiography, photoplethysmography, accelerometry, temperature, respiration, and biochemical indicators collected over extended periods in uncontrolled environments. Missing physiological data are pervasive in such systems because of sensor displacement, battery depletion, motion artifact, communication interruptions, synchronization failures, and patient nonadherence. Conventional imputation methods based on linear assumptions often fail to preserve the complex cross-modal dependencies and temporal dynamics that characterize physiological time series. This paper examines the system-level use of multimodal generative artificial intelligence for missing physiological data imputation. It discusses variational autoencoders, generative adversarial networks, attention-based architectures, and related generative frameworks as components within broader monitoring infrastructures. The analysis emphasizes structural trade-offs among reconstruction fidelity, uncertainty calibration, computational burden, privacy preservation, and clinical interpretability. The paper further addresses governance, regulatory compliance, deployment, robustness, fairness, and lifecycle management. Rather than treating imputation as an isolated statistical task, the paper argues that generative imputation must be embedded within a socio-technical ecosystem that includes data quality assurance, equity auditing, clinical validation, and sustainable operation. The discussion concludes that multimodal generative imputation can materially improve the continuity and analytical value of long-term physiological monitoring, but only when algorithmic performance is accompanied by responsible system integration.

Keywords

missing physiological data; multimodal generative artificial intelligence; long-term health monitoring; data imputation; system architecture; data governance; algorithmic fairness; clinical deployment.

1. Introduction

Long-term health monitoring has shifted from episodic clinical measurements toward continuous, high-frequency observation of physiological states. Wearable devices, home-based monitoring platforms, and ambient sensing systems now generate dense multimodal records that promise earlier detection of deterioration, personalized intervention, and more precise evaluation of treatment effects. However, the inherent complexity of these data is accompanied by substantial missingness. The statistical literature has long distinguished among missing completely at random, missing at random, and missing not at random mechanisms, and this distinction remains foundational for the design of imputation systems [1]. In clinical research, incomplete records can distort risk estimates, reduce statistical power, and undermine the validity of downstream analyses when missingness is handled naively [2]. Multiple imputation by chained equations has become a standard approach for tabular clinical data, but its assumptions are often strained by high-dimensional physiological time series [3].

Generative artificial intelligence has introduced more flexible approaches to missing data imputation. Generative adversarial networks provide a framework for learning complex empirical distributions through adversarial training [4]. Variational autoencoders offer a probabilistic latent-variable formulation that supports principled inference over unobserved values [5]. Foundational work on stochastic backpropagation and approximate inference in deep generative models made these approaches scalable to large datasets [6]. Subsequent models such as GAIN adapted generative adversarial networks directly to missing data imputation [7], while MIWAE extended variational autoencoders to incomplete datasets using importance weighting [8]. Additional work has addressed heterogeneous incomplete data by designing variational autoencoders capable of modeling mixed continuous and discrete modalities [9]. These advances have been supported by public physiological datasets such as MIMIC-III, which provide realistic clinical time series for training and evaluation [10]. Benchmarking studies have shown that deep learning models can extract clinically meaningful patterns from large healthcare datasets, although irregular sampling and missingness remain persistent challenges [11]. Surveys of deep learning for electronic health records similarly emphasize the need for architectures that can integrate asynchronous multimodal observations and propagate uncertainty through imputation [12].

This paper adopts a system-level perspective rather than a purely algorithmic one. It argues that missing physiological data imputation in long-term health monitoring is not only a statistical inference problem but also an infrastructure, governance, and socio-technical design problem. The following sections examine missingness mechanisms, generative architectures, structural trade-offs, governance requirements, deployment considerations, robustness, fairness, and sustainability.

2. The Missing Data Problem in Long-Term Physiological Monitoring

Long-term physiological monitoring differs from inpatient monitoring in several important respects. Data are collected in everyday environments where patient movement, environmental interference, and device handling are highly variable. Scalable deep learning with electronic health records has shown that real-world clinical data contain extensive irregularity and missingness that cannot be eliminated through design alone [13]. A guide to deep learning in healthcare has emphasized that successful deployment requires alignment with clinical workflows, data quality infrastructure, and regulatory constraints [14]. Reviews of deep learning for healthcare also highlight that longitudinal multimodal records are frequently incomplete, with missingness concentrated in periods of clinical interest such as sleep, exertion, or symptomatic episodes [15]. In wearable opto-physiological monitoring,

physical activity introduces considerable motion artifact and signal dropout, and recent work has explored joint attention mechanisms to support robust physiological sensing during such dynamic conditions [16].

Missingness in long-term health monitoring arises from multiple interacting sources. Sensor displacement can interrupt electrocardiographic or photoplethysmographic signals during movement. Battery depletion and charging requirements create structured gaps in overnight or multi-day recordings. Wireless communication failures produce transient losses even when data are acquired locally. Patient adherence varies across time and context, particularly for devices that require active engagement. Synchronization failures across multiple sensors can result in partially observed time points in which some modalities are present while others are absent. These mechanisms are rarely missing completely at random because the probability of missingness is often related to the underlying physiological state, patient behavior, or device performance. For example, a patient experiencing dizziness may remove a monitor or remain motionless in a way that alters signal quality. A generative imputation system must therefore model the joint distribution of observed and missing modalities while remaining cautious about assumptions regarding the missingness mechanism.

The clinical consequences of missing data in long-term monitoring can be significant. Downstream algorithms for arrhythmia detection, fall prediction, metabolic monitoring, or early warning scoring may receive incomplete feature vectors that reduce their reliability. If imputed values are treated as observed measurements, confidence in subsequent decisions may be artificially inflated. Conversely, discarding incomplete segments can remove the very periods during which physiological instability occurs. Multimodal generative imputation offers the potential to retain these segments by reconstructing plausible values from cross-modal context. However, the clinical validity of such reconstructions depends on the fidelity of the learned dependencies and the system's ability to communicate uncertainty to downstream consumers.

3. Multimodal Generative AI Architectures for Imputation

Generative architectures for missing physiological data imputation can be grouped into several broad families. Variational autoencoders learn a probabilistic latent representation of multimodal observations and can infer missing modalities by conditioning on the available ones [5]. The latent variable formulation supports multiple imputation through repeated draws from the approximate posterior, thereby producing a distribution of plausible values rather than a single deterministic estimate. This is valuable in clinical settings where uncertainty quantification matters. Generative adversarial networks, by contrast, learn to produce samples that are difficult to distinguish from observed data through adversarial training [4]. Imputation-specific variants such as GAIN use the generator to fill missing values and the discriminator to assess whether the completed records are statistically convincing [7]. The adversarial approach often yields sharper reconstructions, but it may understate uncertainty because the generator is not explicitly trained to capture the full posterior distribution.

Hybrid architectures combine the probabilistic structure of variational methods with the adversarial pressure of generative adversarial networks. These models can balance reconstruction quality and uncertainty calibration by optimizing both evidence lower bounds and adversarial objectives. Recent work on incomplete heterogeneous data has further shown that modality-specific encoders and decoders can be integrated within a shared latent space to handle mixed continuous and discrete physiological signals [9]. Attention mechanisms have become increasingly important for aligning asynchronous modalities and capturing long-

range temporal dependencies. In particular, joint attention mechanism learning has been applied to opto-physiological monitoring during physical activity to improve robustness against motion-related distortion [16]. Such attention-based fusion is well suited to long-term monitoring because it allows the model to weight informative modalities dynamically according to their reliability.

The choice of architecture has direct implications for system behavior. Autoregressive and recurrent models are natural for sequential data but can be computationally expensive when imputing long records. Transformer-based models can capture dependencies across time and modality more flexibly, but they require substantial memory and careful regularization to avoid overfitting in small clinical cohorts. Diffusion-based generative models have recently emerged as an alternative that can produce high-quality samples, although their iterative sampling process introduces latency that may be problematic in real-time monitoring. At the system level, the imputation module must be selected not only for its statistical accuracy but also for its compatibility with available computational resources, latency requirements, and the need for interpretable outputs.

4. Structural Trade-Offs in System Design

The design of a generative imputation system for long-term health monitoring involves unavoidable structural trade-offs. One central tension is between reconstruction fidelity and uncertainty calibration. A model that produces visually plausible waveforms may systematically underestimate the range of values consistent with the observed context. This can occur when adversarial training encourages the generator to produce high-confidence outputs that match the mode of the data distribution rather than the full posterior. In clinical applications, understated uncertainty is more dangerous than conservative overcoverage because downstream decisions may rely on imputed values as if they were real. Responsible translation of machine learning systems for health care requires explicit attention to such workflow-level risks [17].

A second trade-off concerns centralization versus distribution. Centralized training on aggregated physiological data can yield higher statistical power and easier model management, but it creates privacy risks and may conflict with regulatory constraints. Decentralized approaches such as federated learning can keep raw data within institutions or devices, but they introduce additional complexity in synchronization, communication, and heterogeneity management. A third trade-off involves model expressiveness and computational efficiency. Large transformer-based generative models may achieve superior cross-modal alignment, but they demand significant energy and hardware resources, which may not be available in home-based or low-resource settings. Lightweight models may be deployable on edge devices but may fail to capture rare physiological events.

A fourth trade-off arises between generality and clinical specificity. A generative model trained across many patient populations may learn broad physiological patterns, but it may also obscure subgroup-specific dynamics that matter for equity. Conversely, models trained on narrow populations may perform poorly when transferred to new demographic or clinical contexts. Cross-domain comparisons with industrial process monitoring illustrate the importance of contextual specificity. In manufacturing, missing sensor data can often be imputed using strong physical models of the monitored process. In clinical monitoring, however, individual physiological variability and contextual factors make such deterministic modeling less reliable. The imputation system must therefore be designed as a probabilistic decision-support component rather than as a replacement for clinical judgment.

5. Data Governance, Privacy, and Regulatory Considerations

Multimodal generative imputation operates on highly sensitive physiological data. The system therefore inherits the broader governance obligations associated with health information. Ethical machine learning frameworks in healthcare emphasize data provenance, informed consent, accountability, and the avoidance of secondary harms [18]. A generative model trained on longitudinal physiological records may memorize rare patterns and inadvertently reveal information about individual patients through generated samples. This risk is especially pronounced for high-capacity models trained on small clinical datasets. Governance structures must include access controls, audit trails, and mechanisms for detecting potential privacy leakage.

Regulatory oversight also shapes the deployment of generative imputation. In many jurisdictions, software that influences clinical decision-making may be subject to medical device regulations. An imputation module that feeds a diagnostic algorithm may require validation under the same framework as the diagnostic algorithm itself. This creates a need for documented performance evidence, risk management, and postmarket surveillance. The imputation output must be traceable to the version of the model, the training data, and the operating conditions. Without such traceability, it is difficult to establish accountability when imputed values contribute to an adverse outcome. Data governance must also address the secondary use of imputed records. Once generated values are stored alongside observed measurements, they may be treated as real data in later analyses. Clear provenance markers are needed to distinguish observed, imputed, and partially observed values across the data lifecycle.

The governance framework should also account for patient and provider expectations. Clinicians may be more willing to accept imputed values if the system presents them as estimates with associated confidence intervals rather than as exact measurements. Patients may be more willing to share data if they understand how imputation will be used and what safeguards are in place. Transparency mechanisms such as model cards, data sheets, and audit reports can support informed decision-making. However, these mechanisms must be designed to avoid overwhelming users with technical detail. The system-level challenge is to provide meaningful accountability without transferring an unrealistic burden of interpretation to clinicians or patients.

6. Fairness, Bias, and Socio-Technical Implications

Generative imputation systems are not neutral with respect to social and demographic differences. Bias can enter through the training data, the missingness patterns, the choice of loss functions, or the downstream use of imputed values. An influential study demonstrated racial bias in a widely used health management algorithm because the algorithm used health costs as a proxy for health needs, thereby systematically underestimating illness severity among Black patients [19]. Surveys of bias and fairness in machine learning have cataloged the multiple pathways through which such disparities can propagate [20]. In the context of missing physiological data, fairness concerns arise when missingness is correlated with social determinants of health, device access, or care-seeking behavior. A generative model trained predominantly on data from patients with reliable device access may perform poorly for populations whose data are more frequently incomplete.

The socio-technical implications extend beyond algorithmic bias. Long-term monitoring systems are embedded in clinical workflows, reimbursement structures, and patient-provider

relationships. If imputation improves the apparent completeness of data for some groups more than others, it may create differential access to follow-up care. For example, a model that accurately reconstructs missing cardiac signals for younger, physically active patients but not for older patients with complex comorbidities may inadvertently disadvantage the latter group. Fairness auditing must therefore evaluate imputation performance across demographic, clinical, and socioeconomic subgroups, not only in aggregate. It must also consider the interaction between imputation and downstream decision thresholds. A modest improvement in average reconstruction error may not translate into equitable clinical benefit if downstream algorithms amplify residual differences.

Addressing these issues requires a life-cycle approach to fairness. Data collection protocols should document missingness mechanisms and patient characteristics. Model development should include subgroup-specific evaluation and adversarial testing for bias. Deployment should be accompanied by ongoing monitoring for distribution shifts and differential performance. Feedback from clinicians and patients should inform iterative refinement. This is not solely a technical task. It requires institutional commitment, interdisciplinary collaboration, and regulatory incentives that prioritize equity alongside accuracy.

7. Infrastructure and Deployment Considerations

Deploying multimodal generative imputation in real-world health monitoring requires substantial infrastructure. The system must ingest heterogeneous data streams, align them temporally, detect missingness, run inference, and return imputed values with minimal latency. In acute settings, low-latency imputation may be necessary for real-time alerts. In ambulatory settings, batch processing may be acceptable if the data are used for retrospective review. The infrastructure must also handle device heterogeneity, data format variation, and intermittent connectivity. Edge computing can reduce the need to transmit raw physiological data to central servers, thereby lowering privacy risks and communication costs. However, edge deployment imposes constraints on model size and energy consumption.

Federated learning offers a promising model for training generative imputation systems across institutions without sharing raw patient data [21]. Federated learning in medicine has been demonstrated for multi-institutional collaborations that preserve data locality while enabling shared model improvement [22]. In long-term health monitoring, federated approaches can allow hospitals, home health agencies, and wearable device providers to contribute model updates without exposing individual records. This is particularly relevant for rare physiological events that may not appear in any single institution's dataset. However, federated learning introduces challenges related to heterogeneous data distributions, communication overhead, and security. The global model must be robust to divergent local missingness patterns and demographic compositions. It must also be protected against adversarial updates that could degrade performance or embed bias.

Deployment should also consider the human infrastructure surrounding the model. Clinicians require training to interpret imputed values appropriately. Information technology teams need monitoring tools to detect model drift and operational failures. Patients need clear explanations of how their data are handled and how imputation affects their care. These human factors are often as important as the technical architecture in determining whether a system is used safely and effectively. A phased deployment strategy that begins with retrospective validation and shadow mode operation can help identify unintended consequences before the system influences clinical decisions.

8. Robustness, Reliability, and Clinical Validation

Robustness is a central requirement for generative imputation in long-term health monitoring. The model must perform reliably under sensor noise, motion artifact, missingness bursts, and distribution shift. Unlike static benchmarks, real-world monitoring environments change over time as device firmware updates, patient populations evolve, and clinical practices shift. A model that performs well on a curated evaluation set may fail when confronted with novel device configurations or unusual patient presentations. Reproducibility in machine learning for health research remains a substantial challenge because experimental conditions, data processing pipelines, and evaluation metrics are often underspecified [23]. Clinical validation must therefore go beyond aggregate accuracy to examine failure modes, calibration, and robustness across realistic operating conditions.

Validation should include prospective studies that compare imputed values with concurrently collected reference measurements. For example, in a cardiac monitoring scenario, an imputation system could be evaluated during periods in which both wearable and reference electrocardiographic signals are available, with artificial missingness introduced to mimic realistic dropout patterns. This allows quantitative comparison between imputed and observed values under controlled conditions. In addition, clinical validation should assess whether imputation improves downstream tasks such as event detection, risk stratification, or treatment adjustment. An imputation model that improves waveform similarity but does not improve clinical outcomes may have limited utility. Conversely, a model that modestly improves downstream detection may be clinically valuable even if its reconstructions are not visually perfect.

Reliability engineering is also needed at the system level. The imputation module should be monitored for data quality, missingness rates, model latency, and output plausibility. Automated alerts can flag anomalous imputations or sudden changes in performance. Fallback mechanisms should be defined for cases in which the model is uncertain or the input data are severely degraded. In such cases, the system may abstain from imputing and instead signal that manual review is needed. Abstention is an important but often neglected design option. It preserves safety by avoiding confident but unreliable reconstructions during periods of extreme data loss.

9. Sustainability and Lifecycle Management

Sustainable deployment of generative imputation requires attention to environmental, operational, and institutional dimensions. Training large generative models consumes considerable energy, and frequent retraining can exacerbate this burden. Model compression, distillation, and selective retraining can reduce energy consumption without sacrificing performance. At the same time, the environmental cost of training must be weighed against the potential clinical and operational benefits of more complete physiological records. A lifecycle assessment should consider not only training energy but also inference energy, edge device production, data storage, and eventual model decommissioning.

Operational sustainability involves maintaining model quality over time. Physiological monitoring technologies evolve rapidly, and sensor hardware changes can alter the statistical properties of incoming data. Periodic retraining and recalibration are necessary to prevent performance decay. Data governance policies must specify how long imputed records are retained, how they are labeled, and when they are deleted. Long-term storage of imputed values may create legal and ethical obligations because these values can be mistaken for

observed clinical measurements. Provenance systems must be maintained throughout the data lifecycle to support auditability and correction.

Institutional sustainability depends on the alignment of incentives among technology developers, healthcare providers, regulators, and patients. If generative imputation is used primarily to reduce missingness rates for reporting purposes, it may create a false sense of data completeness. If it is used to support clinical decisions, it must be integrated into workflows in ways that respect clinician time and patient preferences. Sustainable adoption therefore requires ongoing investment in training, evaluation, and governance rather than a one-time algorithmic deployment. The long-term value of generative imputation lies not in producing visually complete datasets but in improving the trustworthiness and clinical utility of long-term physiological monitoring as a whole.

10. Conclusion

Multimodal generative artificial intelligence offers a powerful set of tools for imputing missing physiological data in long-term health monitoring. By learning joint representations across diverse sensor modalities, generative models can reconstruct plausible physiological states even when substantial portions of the record are missing. This capability has the potential to enhance arrhythmia detection, metabolic monitoring, early warning systems, and personalized care. However, the transition from algorithmic promise to clinical practice requires careful attention to structural trade-offs, uncertainty communication, governance, fairness, robustness, and sustainability. Imputation must be embedded within a broader socio-technical infrastructure that supports accountability and preserves patient trust. The continued development of these systems should be guided by interdisciplinary collaboration among machine learning researchers, clinicians, regulatory bodies, and patient communities. Only through such collaboration can generative imputation fulfill its potential as a responsible and durable component of long-term health monitoring.

References

1. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.
2. Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338, b2393.
3. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
5. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *International Conference on Learning Representations*.
6. Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the 31st International Conference on Machine Learning*.

7. Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. *Proceedings of the 35th International Conference on Machine Learning*.
8. Mattei, P.-A., & Frellsen, J. (2019). MIWAE: Deep generative modelling and imputation of incomplete data sets. *Proceedings of the 36th International Conference on Machine Learning*.
9. Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107, 107501.
10. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.
11. Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2018). Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83, 112-134.
12. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589-1604.
13. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1, 18.
14. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29.
15. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236-1246.
16. Zheng, X., Hu, S., Dwyer, V., Barrett, L., & Derakhshani, M. (2026). Joint attention mechanism learning to facilitate opto-physiological monitoring during physical activity. *Biomedical Signal Processing and Control*, 113, 108949.
17. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337-1340.
18. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4, 123-144.
19. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
21. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M.,

- Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119.
22. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milletari, F., Rubin, D. L., Geras, K. J., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 12598.
23. Beam, A. L., Manrai, A. K., & Ghassemi, M. (2020). Challenges to the reproducibility of machine learning in health care. *JAMA*, 323(4), 305-306.