Self-Supervised Physiological Foundation Models for Early Risk Prediction and Secure Human–AI Collaboration in Healthcare

Henrik D. Bush
School of Computing, Clemson University, Clemson, SC, USA.
henrik.bush@clemson.edu

Bjay Srinivasan
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
srinivasan504@uab.edu

Abstract

The transformation of physiological monitoring from episodic clinical encounters to continuous, multimodal data streams opens unprecedented opportunities for early risk prediction and personalized intervention. Self-supervised physiological foundation models, pretrained on vast corpora of electrocardiogram, photoplethysmogram, electroencephalogram, and related biosignals, are poised to become the backbone of next-generation health intelligence systems. This paper presents a system-level analysis of the architectural paradigms, deployment architectures, and governance frameworks necessary to translate such models into trustworthy, secure human–AI collaborations. We examine how contrastive, generative, and masked-reconstruction objectives can capture latent physiological dynamics without costly manual annotations, enabling early detection of cardiovascular, neurological, and metabolic perturbations months before clinical manifestation. The discussion extends beyond model accuracy to encompass the full sociotechnical stack: federated and split learning infrastructures that preserve privacy, edge-cloud orchestration for real-time inference, adversarial resilience against data poisoning and evasion attacks, and continuous monitoring of fairness and calibration drift across heterogeneous populations. Special attention is devoted to the integration of large language model agents in clinical decision pipelines, where secure human-AI collaboration requires adversarial robustness guarantees and transparent uncertainty communication. Through cross-domain comparisons with imaging and natural language foundation models and forward-looking policy analysis, we argue that sustainable, equitable deployment demands institutional mechanisms for auditability, dynamic consent, and algorithmic stewardship. The paper concludes by outlining a research roadmap that couples self-supervised representation learning with system-level safeguards to achieve early risk prediction without compromising patient autonomy or clinical safety.

Keywords

physiological foundation models; self-supervised learning; early risk prediction; human-AI collaboration; adversarial robustness; federated learning; governance.

1. Introduction

The contemporary healthcare landscape is undergoing a fundamental shift from reactive, symptom-driven care toward proactive, data-driven risk stratification. Continuous

physiological monitoring enabled by wearable devices, implantable sensors, and Internet of Medical Things (IoMT) environments generates terabytes of longitudinal time-series data that encode subtle precursors of acute events such as myocardial infarction, stroke, and metabolic decompensation. Deep learning has demonstrated remarkable capacity to extract diagnostic signals from these streams, yet the traditional supervised paradigm is fundamentally misaligned with the scale, heterogeneity, and annotation scarcity inherent in physiological data. Self-supervised learning offers a transformative alternative: by designing pretext tasks that exploit the intrinsic structure of raw signals, large neural architectures can be pretrained on hundreds of thousands of unlabeled records to produce contextualized representations that transfer efficiently to downstream risk prediction tasks.

The concept of foundation models, originally articulated in natural language processing and computer vision, has recently extended into the biomedical domain with notable success in medical imaging and electronic health record analysis. Physiological signals, however, present distinct challenges that require careful consideration of temporality, multi-scale dynamics, sensor modality heterogeneity, and inter-individual variability. The pretraining of a self-supervised physiological foundation model involves architectural decisions about how to encode time, how to handle missing modalities, and how to align representations across varying sampling rates and device characteristics. When carefully executed, such models can serve as a universal feature extractor that enables early risk prediction for diverse conditions, from atrial fibrillation to sepsis, using only fine-tuning on modest labeled cohorts.

Equally critical is the secure collaboration between artificial intelligence and human clinicians. As these models move from research prototypes into clinical workflows, they will be embedded in decision support systems, triage algorithms, and, increasingly, autonomous agentic loops mediated by large language models (LLMs) that interpret physiological data and generate recommendations. This integration amplifies the attack surface: adversaries may craft imperceptible perturbations to biosignals that manipulate model outputs, poison federated training rounds, or exploit LLM agents to produce dangerous advice. The system architecture must therefore incorporate adversarial robustness, privacy preservation, and explainability natively, rather than as afterthoughts, to foster trust and protect patient safety. This paper provides a system-level examination of the design, deployment, and governance of self-supervised physiological foundation models, treating early risk prediction and secure human-AI collaboration as inseparable dimensions of a responsible health intelligence infrastructure.

2. Architectural Paradigms for Physiological Foundation Models

The design space for self-supervised foundation models targeting physiological time-series is considerably broader than that of image or text modalities due to the continuous, multivariate, and non-stationary nature of biosignals. Three dominant architectural families have emerged: transformer-based encoders, convolutional-recurrent hybrids, and state-space sequence models. Transformer architectures, adapted from the vision and language domains, apply self-attention over temporal patches, enabling the model to capture long-range dependencies across heartbeats, respiratory cycles, and neurological rhythms. However, the quadratic complexity of full self-attention with respect to sequence length poses scalability challenges when dealing with days-long ambulatory recordings sampled at hundreds of hertz. Efficient approximations such as Performer and Linformer, or hierarchical approaches that first compress local segments via strided convolutions and then apply global attention, have been proposed to mitigate this burden.

A second paradigm leverages dilated convolutional networks in combination with bidirectional recurrent layers to model multi-scale temporal hierarchies while maintaining linear complexity. This design resonates with biomedical signal processing traditions where wavelet decompositions and filter banks are used to separate frequency bands corresponding to sympathetic and parasympathetic activity. Self-supervised pretraining can be performed using contrastive objectives that pull together augmented views of the same physiological window while pushing apart segments from different subjects or recording conditions. One such approach, CLOCS (Contrastive Learning of Cardiac Signals), demonstrated that representations learned from unlabeled electrocardiogram data could rival supervised models on arrhythmia classification when only a fraction of labels were used [5]. Generative objectives, on the other hand, train the model to reconstruct masked segments of the signal, analogous to masked language modeling in BERT [1] and masked image modeling in MAE [4]. The SIGMA-PPG architecture advances this paradigm by integrating statistical priors of photoplethysmogram morphology into the generative masking process, yielding representations that not only reconstruct the raw waveform but also respect its known physiological constraints [7]. Such inductive biases are essential because pure data-driven reconstruction can ignore subtle pathological deviations that a clinician would immediately recognize, a phenomenon that underscores the importance of domain-informed architecture design.

Multimodal fusion presents an additional architectural challenge. Wearable devices routinely collect multiple channels simultaneously—electrocardiogram, photoplethysmogram, accelerometry, skin temperature—and a foundation model should ideally learn a joint embedding space that aligns these heterogeneous modalities without requiring perfectly synchronized paired data. Contrastive cross-modal objectives, inspired by CLIP [5], can map different physiological views of the same time window into similar representations, enabling downstream zero-shot detection of anomalies across modalities. The architectural choice must also consider deployment constraints: a model pre-trained in a high-performance cloud environment must later be distilled, quantized, or partitioned for execution on resource-constrained edge devices worn on the wrist or implanted subcutaneously. The tension between representation capacity and inference latency demands system-level optimization that spans hardware-aware neural architecture search and federated co-design of model and sensor firmware.

3. Self-Supervised Learning Strategies and Representation Richness

The choice of self-supervised learning objective profoundly shapes the internal representation structure and its downstream utility for early risk prediction. Contrastive frameworks, such as SimCLR [2] adapted to the time-series domain, maximize agreement between differently augmented views of the same signal while minimizing agreement with other signals in the batch. This objective encourages the encoder to become invariant to nuisance transformations—slight timing jitter, amplitude scaling, baseline wander—that do not alter clinical semantics. Yet invariance alone is insufficient; the representation must also preserve fine-grained physiological variations that distinguish normal sinus rhythm from subtle QT prolongation or T-wave alternans. Recent work on variance-invariance-covariance regularization (VICReg) and Barlow Twins demonstrates how to prevent dimensional collapse without explicit negative pairs, an attractive property when minibatch composition may inadvertently juxtapose similar pathological states.

Generative pretraining, exemplified by masked autoencoding, compels the model to develop a deep understanding of signal morphology, phase relationships, and inter-channel dependencies by reconstructing deliberately corrupted input. In the physiological domain, random masking can destroy diagnostically irrelevant noise while forcing the network to infer missing segments from surrounding context, learning a kind of implicit physiological simulator. When statistical priors are injected, as in the aforementioned PPG foundation model [7], the reconstruction is guided to respect pulse wave velocity constraints and the expected relationship between systolic and diastolic phases. This integration of generative modeling with physiological knowledge represents a design principle that can be generalized to other biosignals, including electroencephalogram and capnography.

Beyond purely data-driven objectives, hybrid strategies that combine contrastive and generative losses are gaining traction. A model might, for instance, be trained to reconstruct masked electrocardiogram leads while simultaneously applying a contrastive loss between the latent representation of the original and a time-warped version, thereby reconciling local detail with global invariance. The resulting representations have been shown to carry rich information that linear probes or lightweight decoders can exploit for an array of downstream tasks: arrhythmia classification, ejection fraction estimation, sleep staging, stress detection, and even early indicators of infection as subtle changes in heart rate variability. Importantly, the representation space can be structured to be linearly separable with respect to clinical phenotypes, facilitating interpretability and enabling clinicians to inspect the model's internal organization through projection techniques like UMAP or concept activation vectors.

The scalability of these pretraining strategies is tightly coupled to data governance infrastructure. Large-scale physiological datasets are siloed across hospitals, wearable device manufacturers, and research consortia, each constrained by privacy regulations such as HIPAA and GDPR. Federated self-supervised learning, where multiple institutions jointly train a foundation model without sharing raw data, emerges as a pivotal architectural pattern. Federated optimization of contrastive or masked-reconstruction objectives requires careful handling of non-IID data distributions, communication efficiency, and the risk of gradient leakage that could expose sensitive health information. Secure aggregation protocols and differential privacy mechanisms must be woven into the training pipeline to ensure that the resulting foundation model does not inadvertently memorize individual patient records while still benefiting from the diversity of the federated population.

4. Early Risk Prediction: Infrastructure and Deployment Considerations

Translating a self-supervised physiological foundation model into an operational early risk prediction system involves a complex orchestration of edge computing, cloud inference, model updating, and clinical alerting. A typical deployment scenario envisions a lightweight encoder residing on a wearable or implanted device that streams compressed latent representations, rather than raw signals, to a nearby hub or directly to the cloud. This split-computation architecture trades off privacy, latency, and energy consumption. On-device inference can trigger immediate alerts for critical arrhythmias such as ventricular fibrillation without relying on network availability, while cloud-based downstream predictors can integrate longitudinal trends with electronic health records to compute risk scores for slower-onset conditions like heart failure decompensation.

The dynamic nature of physiological risk necessitates continuous model monitoring and adaptation. Distribution shifts arise from sensor recalibration, seasonal variations, changes in patient medication, and evolving clinical protocols. A foundation model pretrained on a broad

population must therefore be paired with a robust fine-tuning and adaptation regime, such as test-time adaptation or self-supervised domain alignment, that can track a given individual's physiological baseline while guarding against catastrophic forgetting of rare but high-acuity patterns. Early risk prediction for conditions like sudden cardiac death presents a particular challenge, as the event is rare, the premonitory signal may be exceedingly subtle, and the cost of false alarms—both in terms of clinical resources and patient anxiety—is substantial. System architects must design multi-tiered risk escalation pathways where initial low-specificity flags are refined by more expensive, higher-specificity models, possibly involving human review at each step.

The computational sustainability of continuous monitoring infrastructure deserves scrutiny. Running foundation models on millions of edge devices worldwide has an energy footprint that, if unoptimized, could conflict with healthcare's broader environmental sustainability goals. Techniques such as neural architecture search for efficient transformers, weight pruning, quantization, and event-triggered inference—where computation is activated only when a preliminary change-point detector flags an anomaly—can dramatically reduce the carbon cost. The choice between on-device and cloud processing is not merely technical but also ethical: populations in low-resource settings may lack the connectivity or hardware to benefit from cloud-dependent systems, thereby exacerbating health disparities unless offline-capable, compressed models are deliberately designed and deployed.

Interoperability with existing clinical information systems forms another critical infrastructure layer. Early risk predictions must be communicated as structured, standard-compliant messages using HL7 FHIR or equivalent frameworks, ensuring that alerts appear within electronic health record dashboards and are integrated with clinical decision support logic. Without such integration, even the most accurate foundation model will remain disconnected from the sociotechnical workflows that determine whether a prediction leads to a timely intervention. The human-AI interface must present risk scores alongside confidence intervals and supporting evidence, such as a highlighted electrocardiogram strip or a trend of respiration rate over the preceding days, so that clinicians can exercise informed judgment rather than reflexively accept or dismiss the machine's output.

5. Secure Human–AI Collaboration: Robustness, Privacy, and Adversarial Resilience

As physiological foundation models become embedded in clinical care, their interaction with human decision-makers and autonomous agents creates a complex security landscape. The recent proliferation of LLM-based clinical agents that consume structured physiological summaries and generate diagnostic suggestions or treatment recommendations introduces a new vector: an adversarial actor might manipulate a physiological recording in ways imperceptible to human reviewers but sufficient to cause the LLM agent to output dangerous advice. Research on security enhancement methods for adversarial robust LLM agents in medical decision-making tasks has demonstrated that even state-of-the-art models can be vulnerable to carefully crafted prompt injections and context perturbations [6]. In a physiological setting, an attacker could inject subtle, high-frequency noise into a photoplethysmogram recording that shifts the foundation model's embedding toward a region associated with a different pathological state, and the downstream LLM agent, consuming that shifted embedding, might recommend an incorrect medication dosage. System architects must therefore co-design the biosignal encoder and the reasoning agent with adversarial training, certified robustness bounds, and runtime monitoring of consistency across modalities.

Privacy preservation must extend beyond federated training to inference time. When a patient's physiological data traverses multiple nodes—sensor, local gateway, cloud classifier, clinician dashboard, LLM agent—each hop represents a potential leakage point. Techniques such as homomorphic encryption, secure multi-party computation, and on-device differential privacy can ensure that raw or intermediate representations are never exposed in plaintext. In a collaborative human-AI diagnosis scenario, the system may need to provide explanations for its predictions without revealing sensitive training data or model internals that could be used to mount model inversion attacks. The development of differentially private feature attribution methods and zero-knowledge proof-based verification of model behavior are promising directions that align with the broader push toward verifiable and transparent AI in healthcare.

Adversarial robustness in the clinical context cannot be divorced from fairness. Biological signals exhibit significant variation across racial, ethnic, and demographic groups due to differences in skin pigmentation affecting photoplethysmography, chest geometry affecting electrocardiogram lead placement, and underlying prevalence of conditions. An adversary who exploits these correlations can induce model failures that disproportionately harm marginalized groups, turning a security vulnerability into an equity crisis. Defending against such targeted attacks requires not only adversarial training with diverse perturbation models but also population-level stress testing and continuous post-deployment surveillance for emerging disparities. The system must be architected to support robust human-AI collaboration where human overrides and second-opinion protocols are contextually weighted based on model confidence and the known performance profile for the specific demographic segment of the patient at hand.

The operationalization of secure collaboration also demands that the system communicate its own limitations transparently. A foundation model trained predominantly on data from high-income countries may exhibit silent failures when deployed in a tropical disease setting. The human-AI interface should include a continuous honesty signal—a meta-model that estimates the current prediction's reliability based on distributional distance from the pretraining manifold. When uncertainty exceeds a threshold, the system must gracefully escalate to human review, request additional sensor modalities, or actively refuse to provide a recommendation. This refusal capability, while seemingly paradoxical in a decision-support system, is a crucial safety feature that prevents over-reliance and automation bias in clinical workflows.

6. Governance, Fairness, and Policy Implications

The governance of self-supervised physiological foundation models sits at the intersection of medical device regulation, data protection law, and algorithmic accountability regimes. Regulatory agencies such as the U.S. Food and Drug Administration and the European Medicines Agency are grappling with how to evaluate adaptive, continuously learning systems whose behavior evolves with new data. A foundation model pretrained once and then fine-tuned at thousands of sites presents a regulatory conundrum: should each fine-tuned instance be treated as a distinct device, or can the pretrained core receive a foundational certification while downstream adaptations are governed by a site-specific quality management system? The emerging concept of predetermined change control plans, elaborated in the Good Machine Learning Practice guiding principles, offers a pathway, yet significant legal and operational gaps remain [16]. In parallel, data governance frameworks must reconcile the tension between the scale required for effective self-supervised pretraining

and the principle of data minimization enshrined in regulations like GDPR. Federated learning with differential privacy provides a technical substrate, but its deployment at scale requires novel institutional agreements, such as data trust cooperatives that allow patients to collectively negotiate the terms under which their physiological data contribute to foundation model training.

Fairness in physiological foundation models is a multidimensional challenge. Representation bias in the pretraining corpus—for instance, underrepresentation of certain ethnic groups in wearable device studies—can lead to systematically lower accuracy in early risk prediction for those populations. Measurement bias arises because photoplethysmography signal quality degrades on darker skin tones, and electrocardiogram morphology differs with body habitus. Label bias occurs when downstream fine-tuning relies on clinical diagnoses that themselves embed societal inequities in healthcare access. Addressing these entangled biases requires a whole-pipeline approach: pretraining data curation that actively oversamples underrepresented groups, architectural innovations such as domain-adversarial training to disentangle physiological signal from protected attributes, and regulatory mandates for stratified performance reporting before market authorization. Auditing frameworks that perform external, independent evaluation using diverse real-world datasets, akin to financial audits, are essential to ensure that manufacturers' fairness claims are credible and sustained over time [24].

The deployment of foundation models in low- and middle-income countries introduces additional policy considerations. While the global health community has long aspired to leapfrog infrastructure gaps through digital technology, the computational demands of large-scale self-supervised models risk creating a new digital divide where only well-resourced health systems can afford to run and maintain these systems. International policy coordination, supported by organizations like the World Health Organization, could incentivize the development of openly available, computationally efficient foundation model families that can be fine-tuned on modest hardware. Such models would need to be accompanied by open datasets that reflect the genetic, environmental, and pathological diversity of global populations, assembled with informed consent frameworks that are culturally sensitive and not extractive. The long-term sustainability of a global health intelligence infrastructure will depend as much on these sociopolitical arrangements as on raw algorithmic performance.

Human-AI collaboration governance demands that clinicians, patients, and caregivers retain meaningful agency over risk predictions. Informed consent processes for continuous physiological monitoring must evolve from static, one-time agreements to dynamic, granular consent management systems that allow individuals to opt in or out of specific data uses—pretraining, fine-tuning, research, commercial product improvement—without losing access to core clinical services. Algorithmic stewardship committees within healthcare institutions should have the authority to review model performance dashboards, investigate adverse events linked to AI recommendations, and mandate model retraction if safety thresholds are breached. The integration of secure LLM agents into these workflows further raises the question of legal liability: when a clinician follows an LLM agent's advice that was based on an adversarially manipulated physiological embedding, who bears responsibility—the hospital, the model developer, the device manufacturer, or the adversary? Developing a coherent liability framework that does not stifle innovation while ensuring accountability is an urgent policy priority.

7. Conclusion

Self-supervised physiological foundation models represent a paradigm shift with the potential to detect disease risk far earlier than current episodic diagnostic methods, transforming the practice of medicine toward continuous, predictive, and personalized care. The journey from research benchmark to safe, equitable, and secure deployment, however, traverses a landscape of deeply interconnected system-level challenges. Architecturally, the co-design of efficient transformers, informed generative masking, and multimodal fusion with physiological priors will determine whether these models can be sustainably embedded in edge-cloud infrastructures. Strategically, the choice of self-supervised objectives and federated optimization protocols will shape the richness, fairness, and privacy preservation of the representations that downstream clinicians and agents rely upon. Securing human-AI collaboration demands not only adversarial training of biosignal encoders but also robust governance of the LLM agents that increasingly mediate clinical reasoning, a need underscored by ongoing research into security enhancement for such agents [6].

Policy and institutional frameworks must evolve in lockstep with technical innovation. Regulatory agility, international data trusts, algorithmic auditing mandates, and liability reforms are not peripheral concerns but foundational prerequisites for trustworthy health intelligence. Future research should pursue the tight integration of physiological foundation models with verifiable privacy technologies, formal robustness guarantees, and transparent uncertainty communication. Cross-disciplinary collaboration among engineers, clinicians, ethicists, and regulators will be essential to steer these powerful technologies toward outcomes that honor patient autonomy, promote health equity, and protect the integrity of clinical decision-making. By approaching early risk prediction and secure collaboration as a unified system design problem, the research community can help realize a future where physiological intelligence augments, rather than undermines, the human art of healing.

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics.

2. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (pp. 1597–1607). PMLR.

3. Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A., & Gramfort, A. (2021). Uncovering the structure of clinical EEG signals with self-supervised learning. Journal of Neural Engineering, 18(4), 046020.

4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16000–16009). IEEE.

5. Kiyasseh, D., Zhu, T., & Clifton, D. A. (2021). CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In Advances in Neural Information Processing Systems (NeurIPS 2021).

6. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

7. Guo, Z., Chen, T., Jiao, Y., Pan, Y., Hu, X., & Ferrario, M. (2026). SIGMA-PPG: Statistical-prior Informed Generative Masking Architecture for PPG Foundation Model. arXiv preprint arXiv:2601.21031.

8. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

9. Rieke, N., Hancox, J., Li, W., Milletarì, 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.

10. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

11. 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.

12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS 2017).

13. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318). ACM.

14. Chen, M., Hao, Y., Li, Y., Lai, C.-F., & Wu, D. (2018). Edge computing for real-time health monitoring: A survey. IEEE Communications Surveys & Tutorials, 20(4), 2915–2948.

15. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics.

16. U.S. Food and Drug Administration, Health Canada, & Medicines and Healthcare products Regulatory Agency. (2021). Good Machine Learning Practice for Medical Device Development: Guiding Principles.

17. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

18. Moor, M., Banerjee, M., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265.

19. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In International Conference on Machine Learning (pp. 1050–1059). PMLR.

20. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54–71.

21. Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. In Advances in Neural Information Processing Systems (NeurIPS 2019).

22. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175–1191). ACM.

23. Kiyasseh, D., Tadesse, G. A., Nguyen, T. N. T., Zhu, T., & Clifton, D. A. (2021). A clinical report of external generalisation from a deep learning model for cardiac arrhythmias. Nature Machine Intelligence, 3(9), 759–767.

24. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). ACM.

25. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.