

Toward Generative and Causally Grounded Digital Twins in Rare Oncology: Design Principles from Practice

Nimish Saxena,

Andrea Dobrindt

GNQ Insilico, CA, USA

Abstract

Digital twins for rare oncology offer significant potential to accelerate drug development and enable personalized treatment strategies. However, the severe scarcity of data for rare diseases poses a fundamental barrier to construction. Current approaches for clinical digital twin development address synthetic patient generation, or outcome prediction in isolation, but not both, and many lack the interpretability required for regulatory and clinical acceptance. Thus, building digital twins that are simultaneously generative, clinically constrained, and interpretable with limited data remains a methodological challenge.

Herein, we present a methodological framework for clinical digital twin development — implemented as Bio Avatar™, GNQ Insilico’s clinical digital twin platform — and report lessons learned from its application to multiple myeloma. Our work decouples synthetic patient generation and outcome prediction, and trains treatment-specific generative models automatically by selecting between deep learning and causal inference models based on each dataset’s characteristics. Patient generation integrates multi-omic genomic and transcriptomic profiles to capture molecular heterogeneity across patient populations. Clinical validity is enforced through constraint-based validation of physiological ranges, temporal consistency, and treatment sequences. For outcome prediction, we implemented a combined approach utilizing quantum-enhanced optimization, causal inference, and pathway-driven modeling informed by transcriptomic signatures that enable counterfactual reasoning with mechanistic interpretability.

In summary, we identified that synthetic patient generation and outcome prediction are fundamentally different challenges requiring architectural separation. From this foundation, several design principles emerged that challenge

common assumptions in digital twin development. Herein, we provide the methodological and validation framework found during empirical testing and identify areas requiring prospective validation and expansion to other rare cancers.

Introduction

A digital twin is a computational representation of an individual patient that integrates multiple data modalities including clinical measurements, genomic and transcriptomic profiles, medical imaging, and longitudinal treatment records to model disease physiology and therapeutic response. Unlike static patient records, digital twins are dynamic models that simulate a patient’s disease trajectory under different treatment scenarios, enabling in silico evaluation of therapeutic interventions before being administered. By modeling individual patient physiology and treatment response, digital twins offer the potential to accelerate drug development, optimize treatment selection, and enable personalized therapeutic strategies without additional risk to patients^{1,2}. Figure 1 illustrates how digital twins serve as the connective layer between real patient data and in silico treatment simulation in our framework.

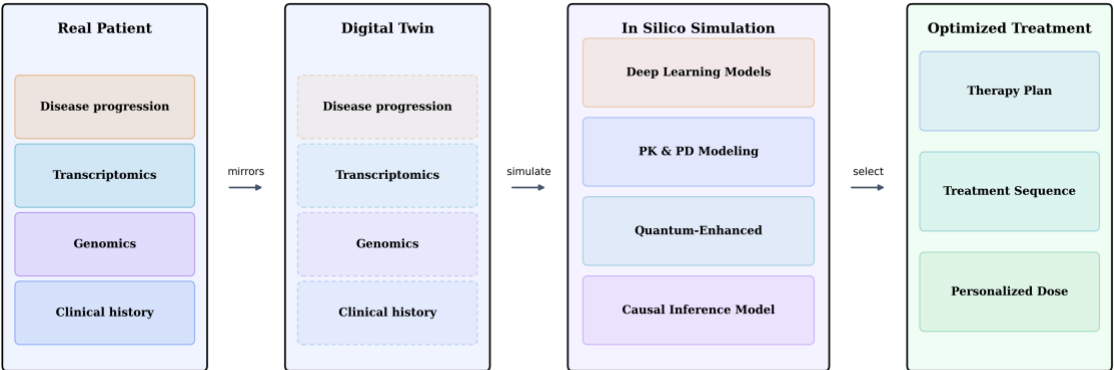


Fig. 1 | Digital Twin In Silico Treatment Framework. Overview of how digital twins serve as the connective layer between real patient data and in silico treatment simulation.

However, rare cancer and oncology , however, the promise of digital twin construction faces fundamental constraint: insufficient of patients. Like most rare cancers, multiple myeloma presents limited cohort sizes, and thus the vast majority of patient-treatment combinations remain unobserved^{3,4}. Current generative methods such as GANs⁵, VAEs⁶, and statistical augmentation were designed for data-rich settings, but the literature offers only limited guidance on applying them under rare disease constraints with limited data. Meanwhile, current digital twin approaches develop

generation and prediction capabilities in isolation, leaving practitioners without integrated frameworks that maintain clinical validity when training data is scarce.

There is a common assumption in digital twin development that models capable of generating realistic patient features are also able to predict treatment outcomes. Recent work applying GANs and VAEs to clinical outcome prediction^{7,8} implicitly assume that learning to reproduce patient distributions entails learning outcome-relevant relationships. However, in a rare oncology setting with limited data, this assumption fails. Generative models interpolate well within training distributions but cannot extrapolate to unobserved scenarios, many of which are of great clinical interest. This limitation is not a shortcoming of any method, and instead, reflects a fundamental boundary of distributional learning under data scarcity.

To address these challenges, we developed Bio Avatar, an integrated digital twin framework for rare multiple myeloma diseases that bridges synthetic patient generation and outcome prediction by taking a combinatorial approach utilizing generative modeling, causal inference, and mechanistic constraints³. The framework was validated against real patient data and exposed architectural principles that current approaches overlooked such as: the necessity of generation-prediction separation, the strength of treatment-specific models over global models, the role of clinical constraints as core infrastructure rather than post-hoc validation, and the value of adaptive model selection that matches method complexity to data availability.

As the field moves toward regulatory acceptance and clinical deployment of digital twins, several architectural questions remain, like how to separate generation from prediction, when to favor simplicity over complexity, and how to enforce clinical validity, to provide the best evidence-based answers. Herein, we provide empirical evidence, drawn from a complete development cycle against real-world rare oncology data, and offer our perspective for others building clinical digital twins under similar constraints.

Approach Overview

Architectural Foundation

The Bio Avatar framework addresses two distinct problems, generating clinically plausible synthetic patients and outcome prediction. Therefore, we treat these two as separate and independent methodological challenges requiring different computational approaches.

This architectural separation reflects a core finding from our development process that generative models excel at producing realistic patient features but cannot reliably predict outcomes for scenarios not represented in training data, as commonly seen in rare oncology treatments. The failure mode of standard synthetic data approaches in rare oncology is illustrated in Fig. 2. By utilizing decoupling while maintaining a shared patient representation, we enabled each component to use methods suited to its specific task whether that was for data-driven approaches for generation or mechanistic approaches for outcome prediction. Following countless iterations, this design framework emerged from our constant efforts to build a unified approach.

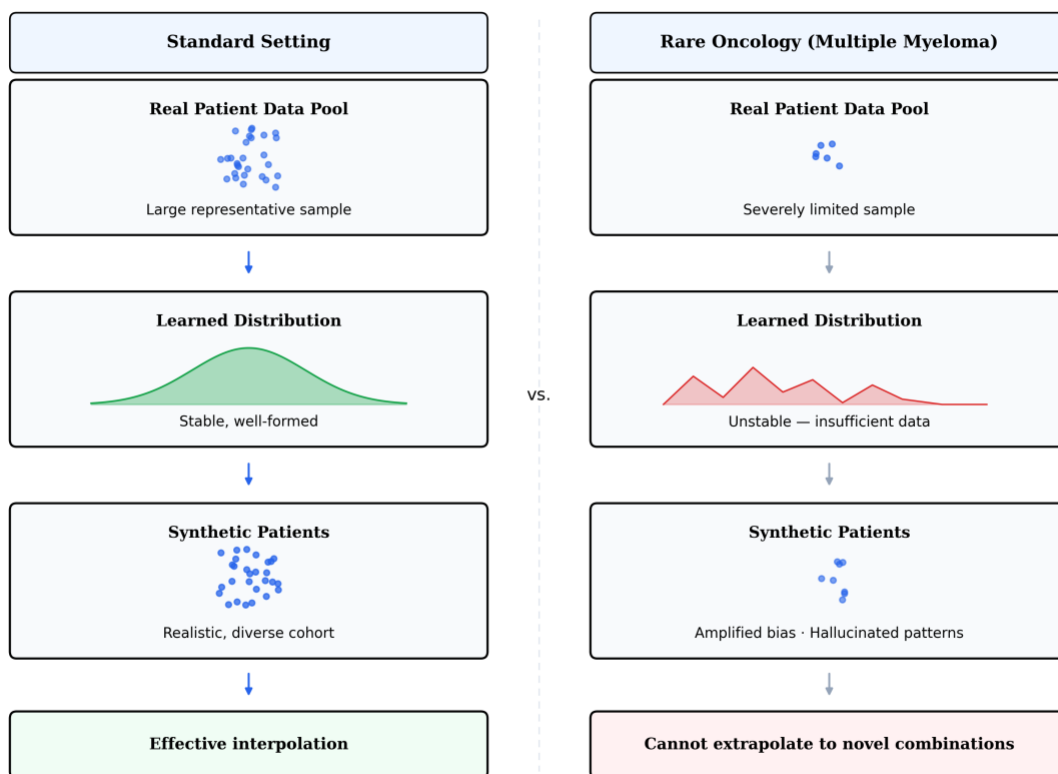


Fig. 2 | Failure of Standard Synthetic Data Generation in Rare Oncology. Standard approaches assume sufficient observations across patient-treatment combinations, a condition rarely met in rare cancers where the majority of combinations remain unobserved.

Generation Pipeline

We trained separate generative models for each treatment regimen rather than a single global model. Separate models preserved clinically meaningful distinctions but introduced a practical constraint, not all treatment regimens have sufficient patients for stable model training.

To address this, we implemented an adaptive model selection strategy that automatically selects between deep learning and statistical generation methods based on dataset characteristics for each treatment regimen. This ensures robust

synthetic patient generation regardless of the underlying sample size. Selection is not based on theoretical considerations, but on empirically observed performance across the development process.

Critically, each treatment regimen produces a uniform number of synthetic patients regardless of its real patient count. This design choice ensures that rare treatments, especially the most clinically interesting in rare oncology, are not marginalized in the synthetic cohort by well-studied regimens that would otherwise dominate purely data-proportional approaches.

Biological Constraint Enforcement

Generative models produce statistically plausible outputs but cannot guarantee medical validity. To overcome this issue, we enforced clinical constraints through post-generation validation, rejecting and regenerating synthetic patients that fail to meet standards derived from clinical knowledge and learned associations in the training data.

Constraint categories include physiological boundaries that no valid patient can violate, temporal consistency requirements for treatment sequencing and disease progression, and treatment-specific patterns that characterize regimens. The specific rules within each category were developed through iterative clinical review, encoding domain expertise that pure statistical methods cannot discover from limited data.

This constraint enforcement operates as architectural infrastructure, not optional post-hoc filtering. Interestingly, we found that the rejection rate during patient generation became diagnostic. High rejection rates signaled that the generative model had drifted from clinical reality and required adjustment. Thus, constraints served as both a validation and as feedback for model development.

Outcome Prediction

To predict outcomes of patient-treatment combinations not observed in training data, we routed synthetic patients through mechanistic modeling rather than rely on the generative model's learned correlations. Our pipeline integrates quantum-enhanced optimization, causal inference^{9,10}, and pathway-driven modeling to navigate the high-dimensional parameter spaces inherent to rare oncology treatment selection. The predicted efficacy and safety profiles generated for each selected treatment combination are shown in Fig. 3, mapping predicted outcomes across key biological pathways to inform treatment selection.

This mechanistic foundation provides biological interpretability that purely data-driven predictions lack. Predictions can be interrogated for biological and pharmacological plausibility, a requirement for both clinical acceptance and regulatory alignment under the FDA Modernization Act 2.0¹¹ that encourages computational alternatives to traditional evidence generation while also demanding transparency in how conclusions are reached.



Fig. 3 | Predicted Efficacy and Safety Profiles Across Treatment Combinations. Pathway-level outcome predictions mapping biological mechanisms to clinical endpoints, enabling mechanistically interpretable treatment selection.

Lessons Learned

Generating twins and predicting outcomes are two distinct but related activities

Initially, we attempted to develop a single unified modeling architecture that would generate synthetic patients and predict their treatment outcomes jointly. We aimed to design and produce one elegant model, trained end-to-end on the complete patient journey, however, we encountered a fundamental limitation.

The generative model we employed learned to produce synthetic patients with features matching training data distributions. Generated patients looked clinically realistic with their characteristics within expected ranges and treatment assignments following learned patterns. However, when we examined predicted outcomes for combinations

not present in the training dataset, such as in rare oncology, the generative model's outputs reflected statistical interpolation rather than mechanistic understanding.

This failure was instructive. Generative models only learn correlations, not causal mechanisms. However, oftentimes for rare oncology, the goal is to explore treatments used in only a handful of patients, and so this limitation is disqualifying for outcome prediction. This implication is significant, and informed us that claims that generative models can predict treatment outcomes for novel patient-treatment combinations should be viewed with skepticism,¹² unless accompanied by evidence of performance on truly out-of-distribution scenarios.

We resolved this architecturally by decoupling patient generation from prediction. Generative models handle their purpose well and learn to produce realistic patient features from limited examples. In addition, outcome prediction routes through mechanistic modeling that can reason through counterfactual scenarios the training data never contained. Interestingly, this separation was not our original design and instead emerged from recognizing the boundaries of what data-driven generation can achieve.

Treatment Specificity Matters More Than Model Efficiency

Our first generative approach trained a single model on the entire dataset. This seemed efficient that one model should be used to represent all treatments and maximize use of available data. However, synthetic patients generated this way exhibited clinically implausible patterns including: toxicity profiles characteristic of one drug class appearing with mechanistically different treatments and mixed dosing patterns across regimens that would never be combined in practice. This pattern collapse resulting from global model training is shown in Fig. 4, where treatment-specific toxicity and dosing signals are lost due to cross-treatment averaging.

Unfortunately, the global model learned to average patterns across treatments, not treatment-specific realities. For clinical validity, this was not acceptable.

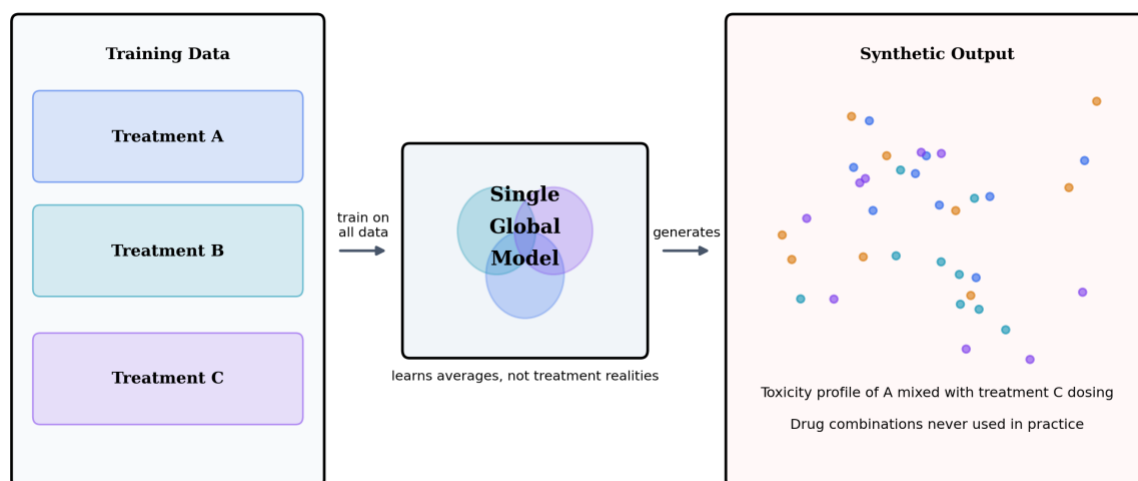


Fig. 4 | Pattern Collapse Under Global Model Training. Treatment-specific toxicity and dosing signals are lost due to cross-treatment averaging when a single unified model is trained across all regimens.

Following the finding that one model cannot do it all, we then pivoted and trained separate models for each individual treatment. This preserved treatment specificity but created a new problem, that many treatment regimens had been used in only a small number of patients. Model training became unstable and outputs showed signs of overfitting or mode collapse.

We were able to resolve this issue by instituting treatment hierarchies. Individual treatments were grouped into parent categories based on clinical similarity, mechanism of action, drug class, and treatment intent. Using this strategy, we increased per-category sample sizes to levels where training became stable and maintained clinically meaningful distinctions. Grouping samples required clinical judgment because sample sizes were too small for data-driven clustering to identify robust patterns.

Fundamentally, this experience exposed a tension inherent to rare disease modeling, the need for treatment specificity conflicts with the data requirements of modern machine learning. However, hierarchical groupings provided a middle ground, though it required domain expertise to define clinically coherent categories rather than merely statistically convenient ones. Therefore, practitioners should expect to iterate between clinical review and model performance when establishing these groupings.

Constraints Are Non-Negotiable

Early synthetic patients passed statistical validation and their feature distributions matched training data, yet during clinical review we identified immediate problems. Laboratory values that never co-occur in real patients suddenly appeared together. Treatment sequences violated temporal logic and physiologically implausible combinations emerged from models that had learned marginal distributions without capturing their dependencies.

To solve this problem, we implemented constraint enforcement through post-generation validation. Synthetic patients failing clinical rules were rejected and subsequently regenerated. Interestingly, the rejection rate itself became a diagnostic. High rejection rates signaled the generative model had drifted from clinical reality and required adjustment.

We found that three categories of constraints proved essential for our analysis. First, physiological boundaries should be implemented to enforce hard limits from clinical standards with values that no valid patient can exceed regardless of context. Second, temporal consistency should be enacted to ensure that treatment sequences, disease progression, and clinical events follow medically plausible trajectories. Third, treatment associations should capture patterns specific to particular regimens, learned from training data and validated against clinical knowledge.

However, regardless of sophistication, purely statistical generation cannot guarantee medical validity. Therefore, constraints derived from clinical expertise are essential infrastructure, not optional refinement. This lesson has broader implications, and illustrates that the literature's focus on synthetic data and distributional fidelity^{13,14} misses a key requirement that clinical applications cannot ignore.

Adaptive Selection Outperforms Fixed Methodology

Not all treatment regimens have sufficient data to use the same modeling approach. Thus, we implemented an adaptive selection approach for each treatment category, and then let the framework select between deep learning and statistical methods based on the original dataset's characteristics.

Deep learning methods were selected only when sufficient training data existed to reliably learn patient distributions. Conversely, statistical methods were used when deep learning would be unstable due to insufficient data, and would produce unreliable outputs, fail to capture variance, or collapse to limited modes.

The selection threshold was determined empirically rather than theoretically. When thresholds dipped below a certain sample size, deep learning methods consistently underperformed simpler statistical approaches that made fewer assumptions about the underlying data structure. Above that threshold, deep learning captured complexities that statistical methods could not represent.

Interestingly, this finding challenges the assumption that more sophisticated methods and models are universally superior. Literature on machine learning implicitly assumes that given sufficient engineering effort, deep learning will outperform simpler approaches¹⁵. However, our results suggest this assumption fails in a rare disease context, and not due to a temporary limitation awaiting better architectures, but rather as a reflection of fundamental mismatch properties between methods optimized for data-rich settings and the realities of rare disease data availability. Therefore, practitioners building rare disease digital twins should not default to deep learning, and instead should match method to context.

Discussion

Modern digital twin methodologies in oncology can be classified into two broad categories: predictive digital twins that forecast disease progression or treatment response from patient-specific models^{16,17}, and generative approaches that create synthetic patient populations to augment sparse datasets^{18,19}. Herein, our framework bridges the two categories by utilizing architectural separation rather than unified end-to-end learning. This decoupling allows for generative capability with predictive modeling, while also allowing each component to use methods best suited for its specific task.

Regulatory Alignment and Interpretability

The regulatory landscape increasingly favors computational approaches²⁰ that can demonstrate interpretability. The FDA has issued guidance on externally controlled trials²¹ and explicitly encourages innovative trial designs, creating pathways for AI-driven methods applied to drug development. However, regulatory acceptance demands transparency on how predictions are generated and why they should be trusted

Black-box models can be highly accurate in aggregate metrics, but face major adoption barriers when individual predictions cannot be explained in terms clinicians and regulators understand²². Mechanistic approaches address this

requirement by design, and predictions emerge from modeled causal relationships including drug mechanisms, pathway effects, and physiological responses that can all be interrogated for plausibility. However, when a prediction fails, it can be attributed to that specific model's components and architecture. Thus, our framework was built with this regulatory reality in mind, not retrofitted to meet it after development.

Limitations

During this study, we encountered several limitations that warrant acknowledgment. First, our framework was developed and tested in a single disease context of multiple myeloma. Whether the architectural principles transfer to other rare cancers or rare diseases more broadly. remains to be validated.

Second, while interpretable, our mechanistic modeling for outcome prediction remains a computational abstraction. It encodes relationships derived from clinical knowledge and pharmacological understanding, but does not simulate biology at the molecular level. Together, integration with experimental validation systems including organ-on-chip platforms^{23,24} organoid models, and other in vitro models will strengthen the biological grounding of predictions.

Third, prospective validation against patient outcomes has not yet been conducted. Our current validation relies on retrospective assessment against held-out data and clinically plausible review. Furthermore, demonstrating framework predictions improve clinical decision-making or trial design efficiency requires prospective studies that we have not performed.

Fourth, while we report the successes and failures during the development process, we do not claim that our solutions are optimal. Perhaps, other architectural choices may prove superior in a different disease context. Thus, we offer our experience as a single data point, and not as definitive methodology.

Future Directions

In the short-term, work should focus on prospective assessment in multiple myeloma and systematic evaluation of transferability to other rare cancers. In the long-term, integration with biological validation systems including organ-on-chip^{23,24}, patient-derived organoids^{25,26}, or other platforms represents an important frontier, creating validation pathways that neither computational nor experimental approaches can offer alone. In conclusion, as regulatory

frameworks for computational evidence rapidly change and evolve, active engagement with regulators will be essential to demonstrate alignment between our framework’s principles and emerging regulatory expectations.

Conclusion

Herein, we describe our findings that synthetic patient generation and outcome prediction are fundamentally different problems that benefit from architectural separation. Generative models produce clinically constrained synthetic patients and mechanistic modeling enables counterfactual reasoning about treatment outcomes. This separation is not a design choice, but a reflection of a fundamental limitation that generative approaches cannot extrapolate beyond what they have observed.

From this foundation, four main design principles emerged: treatment-specific generation outperforms global approaches, clinical constraints are architectural infrastructure not post-hoc validation, adaptive model selection outperforms fixed methodology, and pure data-driven methods cannot substitute for mechanistic reasoning related to unobserved scenarios.

In reality, rare oncology patients cannot wait for the perfect dataset or cohort. Every year, treatment decisions are made with incomplete evidence, and due to bottlenecks, clinical trials struggle to enroll sufficient participants, leading many promising therapies to go untested due to lack of statistical power. Therefore, methods that can generate valid synthetic patients and reason about unobserved treatment scenarios will help close this gap. We offer this methodology framework and our findings as one step closer toward that goal.

References

1. Björnsson, B. et al. Digital twins to personalize medicine. *Genome Medicine* **12**, 4 (2020).
2. Corral-Acero, J. et al. The ‘Digital Twin’ to enable the vision of precision cardiology. *European Heart Journal* **41**, 4556–4564 (2020).
3. Kumar, S. K. et al. Multiple myeloma. *Nature Reviews Disease Primers* **3**, 17046 (2017).
4. Rajkumar, S. V. Multiple myeloma: 2022 update on diagnosis, risk stratification, and management. *American Journal of Hematology* **97**, 1086–1107 (2022).
5. Goodfellow, I. et al. Generative adversarial nets. *Advances in Neural Information Processing Systems* **27**, 2672–2680 (2014).

6. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)* (2014).
7. Yoon, J., Jordon, J. & van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. *International Conference on Learning Representations* (2018).
8. Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R. & Welling, M. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems* **30**, 6446–6456 (2017).
9. Prosperi, M. et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* **2**, 369–375 (2020).
10. Shalit, U., Johansson, F. D. & Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. *Proceedings of the 34th International Conference on Machine Learning* **70**, 3076–3085 (2017).
11. FDA Modernization Act 2.0. Public Law 117-328, Division FF, Subtitle B, Section 3209, 117th Congress (2022).
12. Curth, A. & van der Schaar, M. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems* **34**, 15883–15894 (2021).
13. El Emam, K., Mosquera, L. & Bass, J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *Journal of Medical Internet Research* **22**, e23139 (2020).
14. Bellovin, S. M., Dutta, P. K. & Reitinger, N. Privacy and synthetic datasets. *Stanford Technology Law Review* **22**, 1–52 (2019).
15. Wainberg, M., Merico, D., Delong, A. & Frey, B. J. Deep learning in biomedicine. *Nature Biotechnology* **36**, 829–838 (2018).
16. Niederer, S. A., Sacks, M. S., Girolami, M. & Willcox, K. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science* **1**, 313–320 (2021).
17. Voigt, I. et al. Digital twins for multiple sclerosis. *Frontiers in Immunology* **12**, 669811 (2021).
18. Goncalves, A. et al. Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology* **20**, 108 (2020).
19. Yale, A. et al. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* **416**, 244–255 (2020).
20. U.S. Food and Drug Administration. Considerations for the design and conduct of externally controlled trials for drug and biological products: draft guidance for industry. FDA, Silver Spring, MD (2023).
21. Thorlund, K. et al. Synthetic and external controls in clinical trials – a primer for researchers. *Clinical Epidemiology* **12**, 457–467 (2020).
22. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).

23. Ingber, D. E. Human organs-on-chips for disease modelling, drug development and personalized medicine. *Nature Reviews Genetics* **23**, 467–491 (2022).
24. Low, L. A. & Tagle, D. A. Tissue chips to aid drug development and modeling for rare diseases. *Expert Opinion on Orphan Drugs* **4**, 1113–1121 (2016).
25. Drost, J. & Clevers, H. Organoids in cancer research. *Nature Reviews Cancer* **18**, 407–418 (2018).
26. Vlachogiannis, G. et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. *Science* **359**, 920–926 (2018).