npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01447-y

Survey and perspective on verification, validation, and uncertainty quantification of digital twins for precision medicine

Kaan Sel[1], Andrea Hawkins-Daarud[2], Anirban Chaudhuri[3], Deen Osman[4], Ahmad Bahai[5], David Paydarfar[3,6], Karen Willcox[3], Caroline Chung[2,7] & Roozbeh Jafari[1,4,8] ✉

Digital twins in precision medicine provide tailored health recommendations by simulating patient-specific trajectories and interventions. We examine the critical role of Verification, Validation, and Uncertainty Quantification (VVUQ) for digital twins in ensuring safety and efficacy, with examples in cardiology and oncology. We highlight challenges and opportunities for developing personalized trial methodologies, validation metrics, and standardizing VVUQ processes. VVUQ frameworks are essential for integrating digital twins into clinical practice.

Precision medicine focuses on tailoring health delivery to individuals' unique physiological and disease-related needs based on their unique characteristics and circumstances. Advances in technology and engineering towards personalization have enabled remarkable discoveries across biology, from DNA sequencing and mapping the human genome to identifying specific immune cell types and monitoring electrical activity in the heart. These innovations have significantly influenced medical practice and expanded treatment options for physicians and patients. However, truly leveraging these advances towards personalized medicine has remained elusive. Currently, recommendations for whether a patient should take a drug at a specific dosage or use a medical device are generally based on how their personal characteristics (e.g., sex, age) and disease traits (e.g., genetic mutations, heart rhythm patterns) compare to the population-level data from clinical trials. Given the vast amount of patient data now available, some of which is in real-time, and the significant advances in computational science, this approach is increasingly unsatisfactory. The framework for digital twins, with its capacity for a unique understanding of an individual's current health status as well as predicting their future health trajectories, aligns with the goals of precision medicine. Yet, building accurate digital twins for individuals' health requires confidence in the extracted level of personalized information and the underlying predictive models. Verification, validation, and uncertainty quantification (VVUQ) will be essential to ensuring the reliability of healthcare digital twins and building trust in their clinical application.

Over the past decade, digital twins have gained momentum as computers and algorithms have progressed in their ability to handle large-scale simulations and data processing. The definition of a digital twin varies, ranging from simple virtual representations to models requiring regular, real-time updates based on incoming data[1]. In 2023, the National Academies of Sciences, Engineering, and Medicine (NASEM) published a report to focus the conversation on digital twins and identify critical research gaps[2]. This report put forth a particular definition of a digital twin that we will adopt here: *A digital twin is a set of virtual information constructs that mimics the structure, context, and behavior of a natural, engineered, or social system (or system-of-systems), is dynamically updated with data from its physical counterpart, has a predictive capability, and informs decisions that realize value. The bidirectional interaction between the virtual and the physical is central to the digital twin.* This definition highlights that the concept of digital twins goes beyond only building computational models – they must also be tools for informed decision-making. In precision medicine, this involves creating computational models tailored to individuals' unique physiological characteristics and lifestyle behaviors, enabling precise health assessments, accurate diagnoses, personalized treatment strategies, and the simulation of various scenarios to predict health outcomes to improve individualized care (see Fig. 1). Trust in the underlying processes is critical when dealing with patient health and will influence the acceptance of digital twins by the FDA and healthcare professionals.

Physicians routinely make decisions under conditions of uncertainty due to the combination of incomplete data, inter-patient variability, and evolving medical knowledge. There is growing excitement around the promise of predictive modeling to enable safe, personalized medicine. However, the existence of a digital twin will not eliminate uncertainty. It is crucial that such tools should enhance, rather than undermine, the

[1]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Institute for Data Science in Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [3]Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA. [4]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. [5]Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Neurology, The University of Texas at Austin Dell Medical School, Austin, TX, USA. [7]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [8]Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA, USA. ✉e-mail: rjafari@mit.edu; roozbeh.jafari@ll.mit.edu

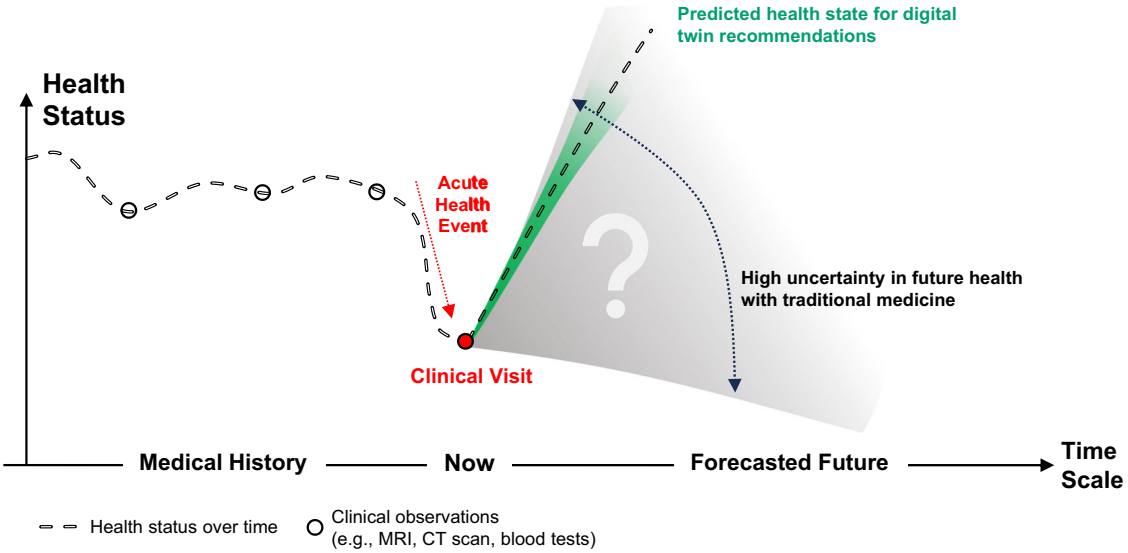

Fig. 1 | Personalized health trajectory predictions using healthcare digital twins for guiding clinical decisions. Digital twins are updated periodically using clinical and ambulatory data. Compared to traditional healthcare, digital twins provide optimized interventions with quantified uncertainty, reducing decision-making uncertainty.

physicians' ability to make clinical decisions. In fact, unquantified uncertainty may prevent physicians from taking appropriate actions or any action due to safety concerns and lack of ability to gauge confidence in the model output. A major research gap identified in the NASEM report was the need for a set of procedures to build credibility and trustworthiness in digital twins. These procedures must address specific questions such as: What is the best way to establish clinical value and trustworthiness of a digital twin, analogous to a clinical trial? Once established, how do we train clinical staff to interact effectively with the tool, understanding its limitations? Once trained, how should caretakers be informed about updates to the digital twin? And can digital twins be updated effectively with new clinical knowledge without compromising trust?

VVUQ was specifically highlighted in the NASEM report as essential for building trust in the use of digital twins for risk-critical applications. Verification, the process of ensuring that a piece of software or a system of software components performs as expected through code solution verification, is a foundational step necessary for establishing trust in software/system solutions. Validation, which tests models for their applicability, aids in understanding the scenarios—such as specific cancer types or treatment regimens—where model predictions can be trusted. Uncertainty quantification (UQ) refers to the formal process of tracking uncertainties throughout model calibration, simulation, and prediction. These uncertainties can be epistemic (e.g., incomplete knowledge of how specific genetic mutations affect a drug's effectiveness) or aleatoric (e.g., natural variabilities not captured by the model). By quantifying these uncertainties, UQ enables the prescription of confidence bounds, which demonstrate the degree of confidence one should have in the predictions. These VVUQ processes have long been conceptually linked in discussions about quality management of software, models, and predictive systems. Numerous methodologies exist for each process, which must be carefully considered in the design of specific applications. As noted in the NASEM report, it is likely that new prediction strategies designed for digital twins in healthcare will necessitate new methodologies for VVUQ.

Digital twins have the potential to significantly impact clinical workflows across various domains, notably in selecting treatment plans based on predicted health trajectories. When paired with proper VVUQ processes, digital twins can become a powerful and reliable tool to simulate interventions and actions that physicians advise. Current practice in medicine involves physicians advising on therapy during a patient's visit; however patient conditions may change. Digital twins, with robust VVUQ support, can transform how physicians consider treatment options with their patients, enabling advanced computational models to inform and perhaps even drive the decision-making and timely administration of interventions at the point of delivery. Therapies and interventions grounded in mechanistic models of digital twins will enable clinicians to make informed decisions based on causal relationships, supported by VVUQ. Mechanistic models emphasize causal inferences, which are critical for helping clinicians identify causation rather than mere correlation. The integration of AI explainability, along with VVUQ and mechanistic models, is likely to create new opportunities for risk assessment by clinicians—opportunities that are not readily available today.

In the meantime, the accompanying need for evolving VVUQ methodologies is becoming increasingly apparent. In this paper, we initially highlight the five components of digital twins and demonstrate the application of the VVUQ processes through cardiology and oncology examples, where promising models have the potential to aid physician decision-making. We then emphasize the role of VVUQ in ensuring the safety, efficacy, and trustworthiness of digital twins in personalized health, particularly for predicting current and future health trajectories based on various interventions and disease models. This paper sets itself apart from previous reviews[3] by focusing on VVUQ for digital twins, highlighting innovations in data integration, model validation, and UQ. It also explores areas for innovation in identifying and quantifying the uncertainties in the clinical decision-making process from a clinician's perspective.

The concept of a digital twin emphasizes the importance of continuously updating the digital model based on its physical counterpart, setting it apart from traditional simulation models. These dynamic updates cannot be limited to routine clinical check-ups, as such an approach could fail to capture the complex and evolving nature of human physiology. Fortunately, advances in biosensor technology now allow for real-time data collection, significantly enhancing digital twin models. This includes a range of clinical-grade in vivo and in vitro sensors capable of monitoring diverse biomarkers. We will explore some of these innovative technologies in greater detail later in this work.

Digital Twins for Precision Medicine for Cardiovascular and Oncology

A digital twin has five main components that are shown in Fig. 2 in the context of precision medicine.

Virtual representation

Central to the digital twin is the virtual representation, which may include mechanistic and/or statistical models that simulate human physiological

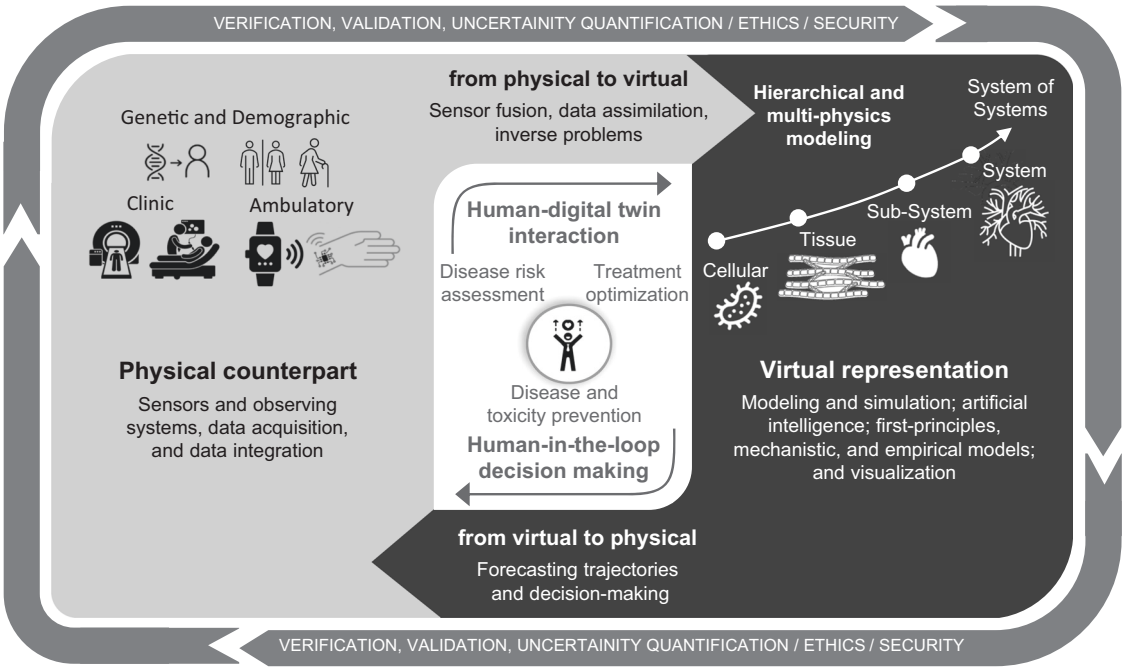

Fig. 2 | Digital twin elements in the context of cardiovascular health. Figure adapted from National Academies of Sciences, Engineering, and Medicine. 2023. *Foundational Research Gaps and Future Directions for Digital Twins*[2]. https://doi.org/10.17226/26894. Reproduced with permission from the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.

phenomena. This virtual representation is tightly coupled with the physical counterpart – i.e., the body and its physiological systems – through continuous observations and flow of information obtained from the physical counterpart. This aspect of digital twins has been extensively studied in healthcare. Mathematical models of cancer and cardiac mechanics, for example, have been in the literature for over 50 years[4]. These models have evolved in complexity ranging from simple ordinary differential equations to agent-based models, stochastic models, and hybrid models, among others. Numerous reviews have examined the variety of these models and their intended use[5–8]. However, these models alone do not constitute a digital twin. As detailed in additional components, these models or virtual representations must be calibrated to their physical counterparts, regularly updated with new information, and allow for bi-directional feedback with a human in the loop. For instance, recent studies present cardiac electrophysiological (EP) models that incorporate CT scans to personalize the anatomical representation, enabling simulations of the heart's electrical behavior at the individual level. These personalized models can aid in diagnosing arrhythmias such as atrial fibrillation (AFib)[9]. Additionally, simulations of myocardial deformation and fluid-structure interactions on an anatomical representation personalized with MRI scans have shown promise for assessing cardiac anomalies, such as atrioventricular valve displacement[10]. Similarly, for oncology, there have been numerous models published looking at predicting overall growth and response to therapy with the aim of identifying response and selecting effective therapies[11].

As the virtual representation evolves with new data, the VVUQ practices must also develop to maintain models' integrity, accuracy, and reliability. While many of these models are run via computer simulations and likely involve some level of code verification, formal verification is seldom discussed in publications[12,13]. Verification for computational models is emphasized in several key reports[14] to ensure applicability, reliability, and robustness of related codes and predictions. Verification involves proving that the coded algorithms correctly solve the intended mathematical model, supported by software quality engineering (SQE) practices[15], and solution verification, which assesses the convergence of mathematical model discretization, typically involving partial differential equations (PDEs)[16]. Validation, as a crucial component alongside verification, assesses how accurately model predictions represent the real world. Publications generally expect models to undergo validation tests, unless they are explicitly focused on hypothesis generation. Despite significant efforts in verification and validation, the validation of models in the context of digital twins raises new challenges. Given the continuous updates and bidirectional data flow, the question arises: how frequently should a digital twin be re-validated to ensure its ongoing accuracy? The dynamic nature of digital twins requires more flexible and iterative temporal validation approaches compared to traditional modeling. It also amplifies the importance of UQ, as each new piece of data can lead to different levels of uncertainties in model predictions. UQ is often lacking in modeling work. Bayesian methods have been used to quantify anatomical uncertainties from clinical data, for example, assessing the impact of MRI data artifacts on the predictive capabilities of electrophysiology simulations[17]. Ultimately, the models selected for use in a digital twin must be 'fit-for-purpose', capable of reproducing the desired quantity of interest with a certain degree of reliability. Selecting the appropriate model is a research challenge in itself, closely linked to both code verification and model validation.

Physical counterpart

The physical counterpart encompasses all elements that interact with data. This component involves the development and utilization of both novel and traditional clinical and ambulatory sensors, along with systems for data acquisition, processing, and integration. The comprehensive physiological information gathered from the physical counterpart can manifest in various forms, including genetic/omic data (e.g., family history, DNA sequence)[18], demographic data (e.g., age, gender, BMI)[19], clinical images (e.g., magnetic resonance imaging, MRI)[20], medical records[21], and continuous data from innovative wearable[22], ingestible[23], or implantable[24] sensors. Typically, imaging techniques like computed tomography (CT), MRI, and positron emission tomography (PET) are employed to create the 3-D geometry for the virtual representation. Advances in biochips and semiconductors now enable clinical-quality, real-time monitoring of biomarkers across various sensing modalities. Semiconductor-based biochemical sensors, such as Ion-Sensitive Field-Effect Transistor (ISFETs)[25], can detect multiple biomarkers in analytes collected from interstitial fluid or plasma. Additionally, wearable

ultrasonic sensors can monitor blood flow through arteries and provide real-time imaging for echocardiograms.

The VVUQ processes must be adapted as data from the physical counterpart is used to calibrate the virtual representation. The availability, quality, and consistency of this data directly impact the 'fit-for-purpose' design of the digital twin. Unlike traditional models built with curated datasets, digital twins must account for the ongoing availability of data at required intervals, ensuring that the VVUQ processes can adapt to potential variability and maintain model integrity over time. Verification of sensors and observation systems generally involves confirming that the systems are implemented correctly according to specifications or design requirements[15,16]. Changes in hardware, whether for cost reduction or performance enhancement, as well as alterations in data structures, must be verified to maintain integrity with the virtual twin. Regarding validation, limitations such as resolution issues and measurement noise must be identified, as they can lead to inaccuracies in the modeled entity. Additionally, data specific to certain disease states, particularly from the target population, are often limited or unavailable. Consequently, digital twin validation using this available data may not be generalizable to the intended population or specific pathology[26]. Contextualizing the physiological data is crucial for guiding the design and calibration of the virtual representation for specific disease models[3]. Monitoring the uncertainties in the physical counterpart is essential to ensure the quality of data flowing to the virtual twin and to track the accumulation of uncertainties within the end-to-end system. Even with abundant data, uncertainties can arise due to noise, missingness, and variance in sensor measurements. Addressing these uncertainties during the information flow from physical to digital is essential for the accuracy and reliability of digital twin models.

Information flow from physical to digital

The third component of a digital twin involves the transition from physical to digital, which entails harnessing vast amounts of data to build and personalize the virtual representation based on the current state of the physical counterpart (e.g., an individual's physiological state) and to continually update the virtual model to reflect changes in the physical state (e.g., aging) and the environment. Essentially, this component ensures that the digital twin accurately mirrors the physical counterpart. The calibration process can vary widely, ranging from simple statistical regressions and computationally demanding inverse problems to training extensive neural networks, depending on the model and data available. Data processing might be necessary, depending on the model of interest and the data at hand. This step is essential as it is a primary point where errors and uncertainties can be introduced into the system. Measurement devices may not always be accurate, and post-processing tasks like image segmentation can introduce additional layers of variability and uncertainty, potentially introducing observer error or bias if any of these tasks require human input. These uncertainties are often quantified using Bayesian inference algorithms, such as Kalman filters and Monte Carlo methods. For example, Kalman filters have been used for measuring uncertainties in clinical hemodynamic observations (e.g., blood vessel diameter) from imaging data (e.g., MRI) for circulation models[27,28]. Markov chain Monte Carlo methods have been employed for estimating parameter distributions from brain MRI to tailor patient-specific radiotherapy regimens[29]. Similarly, several Monte Carlo-based methods have been proposed to manage uncertainties probabilistically during the calibration of electrophysiology (EP) models, such as identifying ablation targets for atrial fibrillation (AFib) treatment from electrocardiogram data[30] and cardiac mechanics models, like determining patient-specific parameters for estimating stroke volume, ejection fraction, and left-ventricular ejection time from echocardiography and blood pressure data[31]. Verification and validation processes must be continuously implemented as the virtual twin is updated – whether periodically or continuously – to ensure that accumulated uncertainties are properly quantified. However, implementing a UQ framework at the organ level often remains challenging due to the complexity of simulations, which require a substantial number of model evaluations. Additionally, creating maps from observables to computational model outputs might introduce further modeling errors[32].

Information flow from digital to physical

The fourth component of a digital twin is the flow of information from digital to physical, manifesting as actionable predictions such as clinical decision-making[33], health trajectory predictions[34], and optimized treatment plans[29,35]. While making actionable predictions is the fundamental purpose of digital twins, and thus may seem straightforward, the process of conveying these predictions is where complexity escalates and building trust becomes crucial. Uncertainty accumulates at every stage of a digital twin – from model to data uncertainty – but it is at this juncture that uncertainty must be carefully communicated along with the predictions[29]. Rigorous methods for UQ and visualization techniques are vital for facilitating discussions about the adoption of digital twins in healthcare and for fostering trust[36], when clinicians are the ones making decisions. However, in cases where digital twins interact with the physical system in an autonomous manner – such as cardiac defibrillators or closed-loop insulin pumps – the VVUQ aspects become even more critical. Here, it is essential to determine the acceptable level of uncertainty based on the 'fit-for-purpose' clinical application, weighing the potential improvements against the risks of erroneous outputs causing harm. Augmented reality[37,38] or web-based visual interfaces[39] can also be instrumental in visualizing digital twin simulations and fostering effective human-digital twin interactions. Such technologies can assist clinicians in making informed decisions, for example, during surgical procedures[40].

Human in the loop

The final listed component is the human in the loop. Although mentioned last, it is a component that permeates the entire digital twin cycle, especially in the context of healthcare digital twins. Humans are deeply involved in the development and operation of digital twins, including tasks such as model selection and validation, data processing, clinical decision-makers, and user interface development[41,42]. While certain engineering applications of digital twins may operate autonomously, this level of automation is far from feasible in precision medicine. Healthcare digital twins by definition, require human involvement, as predictions must be deliberately evaluated in consideration of their inherent uncertainties and the complex factors that define an optimal outcome[43].

Physicians engage in a cognitive decision-making process that, similar to digital twins, incorporates the evaluation of uncertainties and relies on trusted models; they assess available patient data, engage in an iterative diagnostic process by measuring risks, choose tests, interpret results, and optimize the therapy decisions[44]. This process is seldom straightforward due to the presence of uncertainties, including ambiguous or conflicting data and limited evidence of the patient's clinical characteristics[45]. Given these common complexities and similarities in the decision-making process, the human-centric component of digital twins and the corresponding VVUQ processes must be evidence-based to ensure reliability. Integrating structured qualitative methods, such as Clinical Decision Rules (CDRs) is necessary for achieving a consensus in clinical action[46]. Interpretability is another important aspect. Digital twin predictions must be explainable to gain the trust of clinicians and patients. Utilizing interpretable models is essential for a better understanding of the causative factors leading to a prediction, which is vital for clinical decision-making[47]. It is imperative to involve multiple stakeholders within the digital twin ecosystem – ranging from medical practitioners to insurance providers, regulatory agencies, and the patients themselves.

Limitations, Opportunities, Interdisciplinary Perspectives and Outlook

One of the most promising aspects of digital twins in precision medicine is their potential to support decision-making within clinical workflows. To realize this potential, digital twins must meet the same standards of safety, efficacy, transparency, and trust as any other clinical device or support tool.

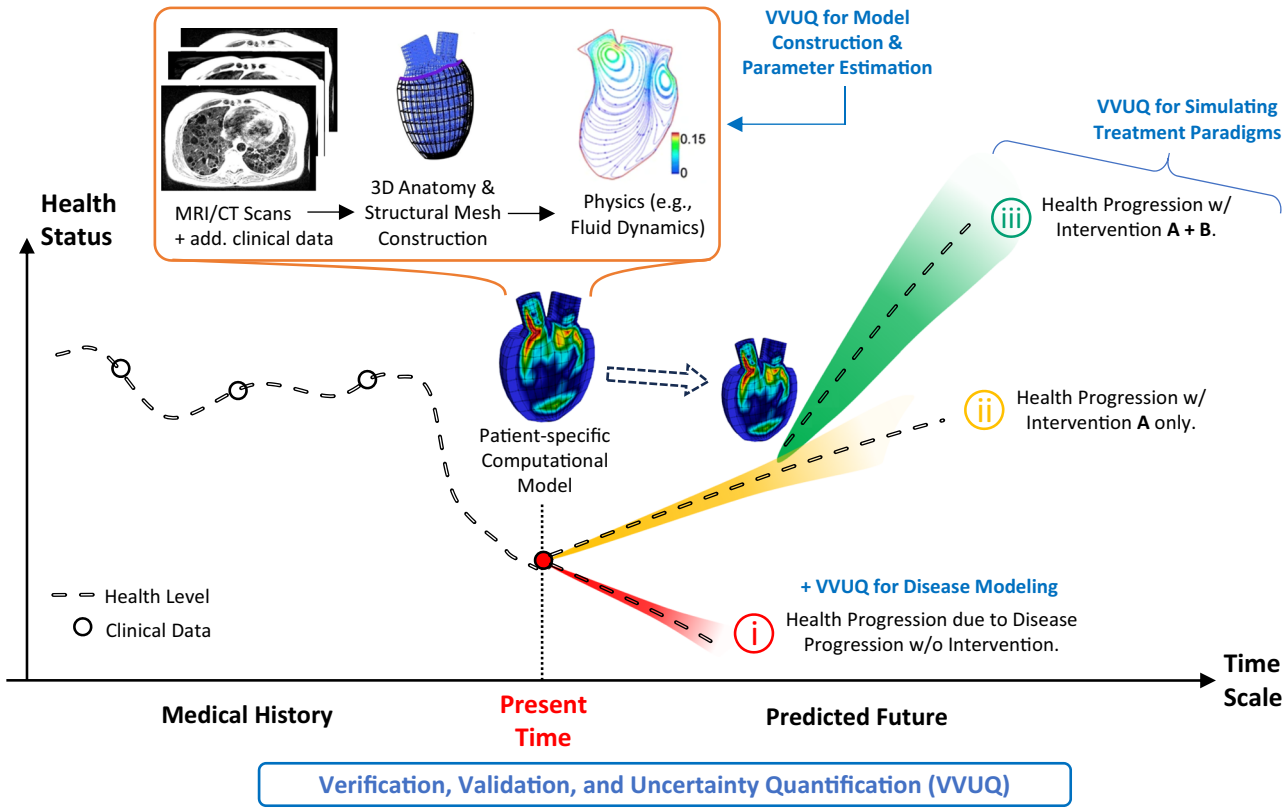

Fig. 3 | Application of the VVUQ processes across different stages of a digital twin's lifecycle, from model construction and parameter estimation to simulating treatment paradigms and disease progression. The integration of patient-specific medical history and clinical data into a computational model enables the prediction of the patient's health trajectory from their current health state. The digital twin also predicts future health progression under different intervention scenarios and disease models (i, ii, iii). The VVUQ processes ensures the trustworthiness and accuracy of the model's updates as new disease and treatment paradigms are incorporated. MRI, 3D structural mesh, and cardiac fluid dynamic simulation images are reproduced from Campbell-Washburn et al. (NIH Image Gallery)[48], and Milosevic et al. [49], respectively under the terms and conditions of the Creative Commons Attribution (CC-BY) License.

VVUQ provides a robust framework that ensures these standards are upheld, facilitating the adoption of digital twins by both clinicians and patients. In *Digital Twins for Precision Medicine for Cardiovascular and Oncology* section, we discussed the application of the VVUQ processes to individual components of digital twins. However, the expectations for digital twins in precision medicine extend beyond predicting health trajectories based on current health states. Digital twins are also expected to facilitate effective, dynamic interactions between the virtual representation, clinicians, and, importantly, patients. Based on the digital twin's projections of a patient's health, patients, in collaboration with their clinicians, can explore various potential interventions and treatment paradigms informed by predicted future health trajectories through disease models. This shared decision-making process enables patients to make informed, personalized healthcare choices, with the guidance of their clinicians. It is essential to ensure that the VVUQ processes remain applicable not only to the digital twin's current state but also to its updates, as it integrates disease progression and treatment paradigms into its predictive modeling, as shown in Fig. 3[48,49]. We acknowledge that validating future predictions is inherently challenging due to the absence of a definitive gold standard for future events. Recent reviews have thoroughly documented the development of digital twins in healthcare, highlighting both their potential and limitations[50–53]. In this manuscript, we emphasize the pivotal role of VVUQ in facilitating the clinical adoption and utility of digital twins. This section discusses key considerations, limitations, and opportunities for digital twins in the context of VVUQ, ensuring safety and efficacy while emphasizing the importance of trustworthiness for successful clinical adoption by both clinicians and patients. Table 1 summarizes key considerations and challenges for developing safe, effective, and trustworthy digital twins for healthcare within the VVUQ framework.

Validation in digital twins extends beyond traditional metrics, requiring adaptive benchmarks that reflect real-world patient variability and aging, along with the clinical objectives

At its core, clinical validation involves comparing model predictions with ground truth data not used in training to determine the utility of these predictions. This comparison can involve various amounts of data, using a single test dataset or cross-validation studies. However, validation tests are not definitive in affirming model accuracy; they merely assess whether the model's predictions achieve a "good enough" agreement with a finite dataset. Consequently, a question for each digital twin is defining what "good enough" means both for its analytical accuracy and clinical utility. The concept of "good enough" should also encompass the broader clinical utility of digital twins within their intended use case. This involves evaluating the accuracy of the model predictions as well as how effectively the outputs of the digital twins translate into improved clinical outcomes. For example, in cardiac digital twins, this may mean assessing whether the predicted outputs effectively guide treatment decisions, optimize surgical planning, or predict adverse events[54]. These outputs must demonstrate relevance, reliability, and robustness under real-world conditions, extending beyond controlled laboratory settings. The core premise of digital twins should be evaluated alongside randomized clinical trials (RCTs), including personalized or N-of-1 trials. While the full potential remains to be explored, robust VVUQ provides a powerful tool for clinicians to design novel treatment strategies that ensure both safety and efficacy[55]. This standard, as mentioned in the NASEM report, should be established based on "fit-for-purpose" principles, considering multiple factors such as intended use and potential consequences. The validation process should also account for potential mismatches between the context in which the model is developed and its intended use. Instances where the model is contraindicated – when

Table 1 | Key Requirements for Ensuring Safety, Efficacy, and Trustworthiness of Digital Twins in Precision Medicine

	Key considerations and challenges in the context of Verification, Validation, Uncertainty Quantification.	Key Attribute
1	Validation assesses whether a model's predictions are "good enough" based on a finite dataset driven from clinical outcomes. Defining what constitutes sufficient accuracy should be driven by the application.	Efficacy & Trustworthiness
2	Validating digital twins due to their individualized nature (N-of-1) require approaches beyond conventional RCTs. The absence of control groups can be addressed through personalized trial methodologies that randomize treatment periods within a patient. This allows for precise evaluation of safety and efficacy while directly linking predictions to clinical outcomes. Additionally, the development of new metrics and statistical methods tailored to personalized trials is essential. Cohort-level validation can complement this by identifying commonalities in data inputs, outputs, and model characteristics, ensuring broader applicability without sacrificing individualization.	Safety, Efficacy, & Trustworthiness
3	While data-driven algorithms can create statistical models for digital twins, it is essential that models remain physiology-aware and mechanistically grounded, preserving the underlying biology and physics. This approach is critical for accurately quantifying uncertainty in predictions and ensuring interpretability.	Trustworthiness
4	Although uncertainty quantification is central to establishing trust, the relative weighting of digital twin predictions with the measured uncertainty should depend on the baseline uncertainty of the clinical scenario and the specific context of its application.	Efficacy & Trustworthiness
5	Digital twins in the clinical setting must balance the trade-off between early intervention with less data and higher uncertainty and delayed intervention with more data and greater certainty but possibly reduced magnitude of impact due to the time delay for action based on the prediction.	Safety, Efficacy, & Trustworthiness
6	Standardizing VVUQ processes is crucial for ensuring interoperability, quality assurance, and risk management in digital twins, similar to clinical trials, while addressing challenges posed by the limited interpretability of statistical models like AI and machine learning.	Safety

the digital twin predictions are "out-of-specifications" for the patient due to the patient's condition or disease state does not match the conditions for which the model was validated – should be identified and avoided. One potential approach to address such "out-of-specifications" is the use of knowledge graphs, which organize and categorize relationships between data features and relevant knowledge bases, ensuring that digital twin predictions align with the clinical knowledge associated with the intended use[56]. An example application of knowledge graphs is in identifying drug-drug and drug-disease interactions before the drug is released for clinical use[57]. The demand for robust data has spurred research into next-generation sensors capable of pervasive physiological monitoring. These sensors are being designed to frequently, if not continuously, capture personalized physiological data across various contexts (e.g., health states, environmental conditions)[58] to inform and calibrate digital twins. The nature of the observational systems and data characteristics may show significant variation based on the medical application that the digital twin is created for. A particular example is digital twins for mental health, where a significant portion of the data is sourced through surveys, questionnaires, and self-reported symptoms of subjective markers such as pain, discomfort, or stress[59,60]. This multi-sourced high-dimensional information requires a different approach for validation and uncertainty quantification when compared to biomarkers of specific diseased states. For example, pain is a subjective symptom that can be associated with specific underlying pathology. The use of digital twins enables sensitivity and causation analysis to uncover the root causes of pain. This approach allows for tracking the progression of pain alongside a diseased state and evaluating the response to treatments. By doing so, personalized outcome measures can be developed, benefiting both patients and physicians. The VVUQ processes should be tailored to rigorously analyze the sources for potential biases, variances, and measurement errors in the observational data to ensure the digital twins establish a robust patient-specific physiological and behavioral profile, including individuals' emotions, perceptions, and behaviors[61]. This will also ensure that the digital twin predictions are validated against both subjective reports and objective measures and the confidence levels in the predictions are calculated, while accounting for the variabilities in the subjective data. Here, structured approaches, such as measurement feedback systems, are proposed for defining valid, reliable, and standardized metrics for validation[61,62]. Additionally, Bayesian approaches may allow for quantification of uncertainties, enabling confidence estimates as new behavioral, psychological and physiological information is obtained and used for updating the digital twin. For example, probabilistic graph models, creating a dynamic Bayesian network, are used for coupling between different data collection sessions for the underlying dynamic physiological and psychological system[59,63].

The individualized nature of digital twins challenges traditional validation metrics, while presenting an opportunity to redefine efficacy evaluation to measure success

Clinical trials for drugs and devices, often termed "validation studies," typically focus on population statistics – such as whether one patient group statistically outperforms another. In contrast, digital twins aim to personalize therapy to optimize outcomes at an individual level, creating a paradigm shift where each personalized digital twin functions as an individualized experiment of N-of-1. This shift requires precise calibration to the individual to ensure accurate predictions, without relying on traditional control groups used in conventional RCTs. An opportunity to address this is to leverage methodologies developed for N-of-1 trials. Unlike RCTs, which randomize groups of patients to treatments, N-of-1 trials randomize treatment periods within a single patient[64]. This approach directly addresses challenges associated with averaging treatment effects across populations, ensuring external validity despite heterogeneity in treatment effects. Furthermore, N-of-1 trials are particularly well-suited for digital twins, enabling precise representations of rare diseases or complex comorbidities when traditional RCTs are not effective. The iterative and individualized nature of these trials accelerates the process of treatment discovery and their implementation, offering a powerful framework for advancing personalized treatment efficacy. Recent work has proposed a computational framework that treats each patient as their own trial, based on four key techniques[65]. First, mathematical models are calibrated with patient-specific data to predict personalized responses. Digital twins then simulate different treatment strategies, refining them using optimal control theory to maximize outcomes. Furthermore, data assimilation continuously integrates new clinical information, allowing the digital twin to adjust predictions and interventions in real-time. This iterative process reduces uncertainty and refines optimized intervention based on up-to-date data. Another opportunity, or in certain circumstances a requirement, when access to personalized data is not possible, is to use a sub-population data both for digital twin calibration and validation. The key is to identify a representative cohort that shares commonalities with the patient-of-interest in terms of the system inputs/outputs, boundary conditions, and model characteristics. Traditionally, this process relies on clinical experts manually defining ranges for data patterns (e.g., blood pressure > 140/90 mmHg) for a specific outcome of interest (e.g., hypertension diagnosis). However, this approach is time-intensive, requires extensive medical knowledge, and often overlooks

confounding variables that influence outcomes. A robust alternative includes leveraging machine learning and artificial intelligence models[66] such as regression[67] and recurrent network[68] large language[69,70] models, to extract representative features and assist in cohort discovery. While these approaches can generalize beyond the N-of-1 paradigm, challenges such as bias amplification, confounding, and causal inference remain critical and require careful assessment[71], as well as the use of statistical methods such as propensity score matching[72] (by calculating and matching the probability of an individual having specific characteristics such as disease state or treatment exposure based on observed covariates) or PROCOVA (Prognostic Covariate Adjustment) methods for additional control[73]. Additionally, as digital twins update to reflect changes in the "aging" physical counterparts, a previously validated version of a digital twin does not necessarily guarantee the validity—such as safety and efficacy—of its subsequent updates. Here, digital twins developed in other industries can offer valuable technical insights. For example, in cyber-physical systems, continuous validation is achieved by tracing the alignment between the digital twin and its real-world counterpart using Needleman-Wunch algorithms[74,75]. This approach could potentially be adapted for precision medicine. Central to these efforts is VVUQ ensuring a tight coupling between virtual and physical representations. The concept of "models of models" can be leveraged in digital twin frameworks, allowing multiple virtual representations to run in parallel and selecting the most suitable model dynamically. Beyond classical VVUQ approaches, incorporating data consistency checks between the models and the physical counterpart, as well as using benchmarks to test edge cases, can further enhance robustness and reliability[76,77]. Addressing these challenges when defining metrics for success will likely take many forms. We anticipate early efforts will still rely on aggregated measures, but as digital twins evolve in sophistication, able to capture many more scenarios, statistical methods may co-evolve to better answer whether a digital twin for a particular individual is valid.

Trustworthiness in digital twins begins with the explainability of decisions, the factors influencing those decisions, and the uncertainties quantified for those decisions

The interpretability of the underlying models is essential for providing transparent decisions and understanding the causations behind predictions. While deep learning and other data-driven algorithms can serve as surrogates for aspects of the virtual representation where purely mechanistic modeling is computationally prohibitive, it is crucial to prioritize interpretable models that integrate biological and physics-based knowledge. One promising approach is to combine the predictive power of artificial intelligence with fundamental physics to develop scientific machine learning (SciML) techniques[78], including physics-informed neural networks (PINNs)[26]. Moreover, reduced-order models, Gaussian process surrogates, and regression (or classification) trees[79] can provide computational efficiency while ensuring that models remain interpretable and their predictions are quantifiably reliable – i.e., UQ and communication to physicians and patients. Additionally, the use of both computational and physical phantoms can serve as benchmarks for validating digital twin predictions and testing UQ frameworks[80,81]. When predictions with varying confidence levels are presented, clinicians need clear thresholds for success and failure based on the associated confidence levels. This helps ensure that healthcare professionals can rely on digital twin predictions in clinical decision-making. Furthermore, ensuring the secure and private transmission of patient data from the physical counterpart to the digital twin remains a critical challenge.

An equally important aspect of trustworthiness concerns the ability of digital twins to execute within reasonable time frames, making the quantifiably reliable decisions usable in clinical settings

This presents a significant challenge, as the accuracy and complexity of the underlying models often increase their computational costs. Managing uncertainty in this context requires multiple evaluations of such costly models, making the deployment of digital twins computationally demanding and potentially delaying decision-making for clinicians. Recent studies have demonstrated that high-fidelity, multi-scale cardiac model simulations with UQ demand extensive computational resources, taking several hours to simulate a single heartbeat and years of approximated execution time on multi-core processors for complete global sensitivity analysis simulations[82]. The same data science strategies that add interpretability can be leveraged to improve computational efficiency. These include surrogate methods[83], scientific machine learning techniques[82], and multi-fidelity methods[84,85]. Reasonable approximations and simplifications of the underlying complex systems may be crucial for achieving convergent predictions with quantifiable uncertainties in actionable time. For instance, in electrophysiology models of cardiac tissue, one potential strategy might involve constructing electric network representations of blocks of cardiac myocytes rather than modeling individual cells[86]. This approach reduces the computational burden of constructing and solving extensive electrical networks that represent the intra- and extra-cellular coupling of every heart cell, while still preserving modeling accuracy by leveraging tissue continuum properties.

A trade-off in the deployment of digital twins for precision medicine lies in the timing of interventions versus the associated uncertainty

Early prediction and simulation of disease trajectories and health progression due to potential future intervention paradigms, inherently come with higher uncertainty due to the limited availability of specific data points early in the disease progression[87]. As time passes and more relevant data becomes available, the uncertainty may decrease[88,89]. This dynamic presents a dilemma: intervening early can potentially allow for more effective disease management but carries the risk of lower certainty in the outcomes. Conversely, waiting until closer to the disease manifestation reduces uncertainty but may limit the effectiveness of the intervention[89,90]. Therefore, determining the optimal decision-making point where the benefits of early intervention outweigh the risks posed by uncertainty is crucial for maximizing the impact of digital twins in clinical practice. This decision is also shaped by the baseline success rates of current clinical approaches and the potential for improving upon them. For example, in cases where existing outcomes are particularly poor, accepting higher uncertainty in an intervention may still be justified, similar to the rationale used in Phase I clinical trials, where higher risks are often deemed acceptable.

Digital twins for precision medicine are expected to provide patient-specific recommendations to clinicians, where clinicians combine their knowledge with the digital twin recommendations for clinical-decision making. Given their crucial roles, digital twins must adhere to FDA regulations intended for clinical decision support (CDS) devices[91]. These regulations should rigorously ensure validity, safety, and efficacy of the digital twin predictions. However, digital twins often process complex multi-dimensional data through models and mechanisms that may lack interpretability, posing challenges for analytical validation and risk assessment[92]. Consequently, the traditional regulatory framework established by the FDA for medical devices is not fully suitable for ensuring the effectiveness and safety of digital twins[93]. In recent years, the FDA has provided guidelines for assessing the credibility of computational models and simulations used in medical devices[94]. However, as digital twins require ongoing updates and VVUQ integration to accurately reflect changes in their physical counterpart with calculated risks, their dynamic nature represents a regulatory challenge that the existing framework does not fully address[95].

While each digital twin will necessitate its own testing scenario, standardizing the VVUQ processes as much as possible is critical for ensuring interoperability, quality assurance, and effective risk management, especially given the high-stakes nature of clinical decision-making often involved with digital twins[53]

Thoughtful standardization will also facilitate communication with audiences familiar with previous digital twins, enabling them to quickly grasp

new concepts. This approach mirrors the standardized phases of clinical trials: Phase 1 focuses on safety with a small group, Phase 2 expands slightly to evaluate both safety and efficacy, and Phase 3 tests for statistical significance in outcomes at the population level. Existing standards and guidelines from computational modeling, simulation, and data sciences offer valuable reference points. For instance, the 2018 Validation and Verification for Medical Devices (V&V 40) standard by the American Society of Mechanical Engineers (ASME) provides a framework for assessing the efficacy of computational models[94,96]. However, there remains a gap in properly standardizing VVUQ implementation and reporting specifically tailored to digital twins. This gap is partly due to the increasing reliance on statistical modeling techniques, such as machine learning and artificial intelligence, which introduce challenges in interpretability and validity compared to traditional mechanistic models. Addressing this gap will require active engagement with medical communities of practice, which play a crucial role in defining consensus standards and establishing parameters for the clinical adoption of digital twins. These communities act as gatekeepers, ensuring that digital twin outputs align with evidence-based standards and thresholds necessary for establishing clinical guidelines. Such consensus-building is critical for integrating digital twins into clinical workflows, as it ensures outputs are actionable and trusted by clinicians. By collaborating with these communities, VVUQ processes can evolve to bridge the gap between technological innovation and practical implementation, fostering widespread adoption and confidence in digital twin technologies. Effective standardization of VVUQ will be essential for integrating digital twins seamlessly into clinical workflows, ensuring their design and continual updates maintain reliability, efficacy, and safety at all times. While the standardization of VVUQ processes is crucial for the successful integration of digital twins into clinical workflows, real-world examples that meet regulatory approval remain limited. One of the most notable cases of validated UQ and regulatory acceptance in digital twin technology is HeartFlow's work in diagnosing and planning treatment for coronary artery disease (CAD)[97]. HeartFlow's platform combines anatomical and physiological data from coronary computed tomography angiography (CCTA) with AI-driven analysis to create patient-specific models. In 2022, HeartFlow became the first company to receive FDA 510(k) clearance for its AI-powered Plaque and RoadMap analyses, in addition to its established physiology analysis based on CCTA[98]. A key factor in HeartFlow's FDA approval was the ability to ensure the safety and efficacy of the proposed technology through formal mathematical modeling and VVUQ. HeartFlow's success underscores the critical role of VVUQ frameworks in achieving regulatory acceptance and integration into clinical practice, ensuring not only safety and efficacy but also practical, real-world application. In addition to the HeartFlow platform, other advanced computational platforms such as Medis QFR and CathWorks FFRangio integrate AI to enhance cardiovascular care. Medis QFR offers a noninvasive assessment of angiography-derived physiological simulations for patients with epicardial artery disease, integrating AI to ensure accurate evaluations of the coronary arteries. Similarly, CathWorks employs its AI-driven FFRangio technology to calculate functional flow reserve values from coronary angiogram images, guiding clinical decision making[99,100]. Collectively, these platforms align with the goals of the Living Heart Project, a collaborative initiative to advance patient-specific cardiovascular simulations. This project highlights the shared commitment to leveraging advanced modeling and AI to improve diagnosis, risk assessment, and treatment planning[101]. There are also noteworthy examples of complex purely data driven AI models for clinical diagnostics, where the performance depends heavily on the quality and quantity of input data. Meeting these requirements can result in models that potentially match, or in certain instances, outperform clinician assessments. For example, AI models for diagnosing mammograms reduce the false positive rates[102], enhance sensitivity for classifying patients with breast cancer[103], and can detect diabetic retinopathy[104]. AI algorithms can also help reduce errors in the clinical decision making while providing faster diagnostic results by analyzing subtle patterns in historical and current medical data[105]. However, human oversight is essential to assess the effectiveness of the model's diagnosis for at-risk patients, along with ensuring access to large volumes of high-quality data. Challenges such as data privacy, patient data security, and model hallucination, where models generate false yet plausible results with high confidence, remain significant concerns[106,107]. Further obstacles arise from the conditions under which AI models are validated. For instance, model accuracies are often tested in controlled settings that do not reflect clinical environments, with limited external validation and few real-world studies conducted[108,109]. Due to lack of interpretability in most AI algorithms, the rationale behind AI-driven diagnoses or recommendations is often unclear leading to mistrust among clinicians[110]. Addressing these limitations, which are also essential to supporting VVUQ for digital twins, could involve enhancing literacy of the models and trust among clinicians, incorporating more diverse datasets, and establishing standardized testing and interpretability procedures to properly verify and validate the models[111].

An innovative direction for regulatory bodies could involve utilizing digital twins to conduct virtual clinical trials before proceeding to Phase 1 trials

Currently, clinical trials rely heavily on experiments with animal or human subjects, restricting the scope of interventions and often limited to small cohorts that may not fully represent the target population. By integrating digital twins in the clinical trial process, regulatory bodies like the FDA could conduct virtual (pre-)clinical trials on simulated patient cohorts. This approach would significantly accelerate the trial process and reduce risks by enabling the preliminary evaluation of interventions on a diverse and extensive virtual population. For instance, AI-driven digital twin models incorporating PROCOVA technique – used for reducing the variances included for the population selection – qualified by the European Medicines Agency (EMA) demonstrates augmentation of the control groups in clinical trials, reducing the need for large placebo groups while maintaining scientific validity[73]. Moreover, these virtual trials could swiftly identify potential failure candidates and extensively assess the safety of drugs or medical devices, enhancing overall efficiency and effectiveness in drug and device development.

While the VVUQ processes are fundamental for ensuring the safety, efficacy, and trustworthiness of digital twins in precision medicine, several limitations must be acknowledged

These include: i) complexity versus time requirements: The VVUQ processes, particularly the UQ component, are computationally intensive. For example, conducting comprehensive sensitivity analyses across extensive parameter spaces demands significant computational resources, which may not be feasible within tight time constraints for the target clinical application[82]. ii) Data demand: The validation process for patient-specific digital twin necessitates the availability of control data collected from the patient. In most clinical settings, such as for rare conditions or certain acute events (e.g., heart failure), such representative data is not available. In certain cases, a systematic broader patient cohort analysis may address this shortcoming by identifying representative sub-population control groups, leveraging commonalities in data inputs, model functionalities, and outcomes[112]. Additionally, personalized trials (N-of-1) could serve as a valuable strategy to generate personalized control data in cases where broader cohort analyses are infeasible[55]. Digital twin technology with robust VVUQ will enable the realization of personalized trials. Strong VVUQ ensures predictive safety and efficacy in testing treatment strategies. With personalized models, RCTs with group- or phenotype-based designs can shift to individualized trial approaches, allowing for safer, more effective, and faster evaluation of treatment strategies. iii) Scalability: the need for digital twins to be 'fit-for-purpose' for specific clinical applications may complicate the scalability and standardization of the VVUQ process, as each unique application may require unique adjustments to the VVUQ process. iv) Interpretability and usability: The complexity of the VVUQ processes may pose interpretability challenges for patients and clinicians, making it difficult to be utilized in the clinical workflow for assisting clinical decisions.

Several tools have been proposed to communicate quantified uncertainties and risks to patients and clinicians, leveraging statistical techniques such as the Kaplan-Meier curves used for optimizing patient-specific radiotherapy regimens[29], or explainable artificial intelligence (XAI) methods designed to visualize head and neck cancer locations and spread[113], v) Dynamic nature of digital twins causing operational burden: As digital twins evolve alongside their physical counterparts, integrating a continuous VVUQ processes to accommodate updates can impose significant operational challenges. This may result in periods where the digital twins operate under increased uncertainty or without proper validation. vi) Ethical concerns: Rigorously applying VVUQ processes requires continual access to highly personalized and potentially sensitive data, such as medical records and genomic information. This raises significant ethical concerns, including issues related to patient consent and data privacy, which can complicate the implementation of digital twins with integrated VVUQ processes. These challenges highlight the need for ongoing research to refine VVUQ methodologies and address these limitations to enhance the practicality, applicability, and privacy of digital twins in clinical settings.

Incorporating patient-level health data into digital twins for precision health pose several privacy challenges

Particularly, medical records and genetic information used for design, calibration and updates of digital twins contain highly sensitive patient information, where an unauthorized or unconsented access or breach can lead to significant privacy violations. Moreover, combining multi-level personalized data to enhance the accuracy and utility of predictions, along with adding interpretability to model predictions to improve trustworthiness, may inadvertently increase the risk of re-identification. To address these challenges, the VVUQ framework must incorporate robust data governance strategies, protecting patient confidentiality and complying with data protection regulations. Potential solutions include the implementation of advanced encryption techniques, ensuring secure data storage, and establishing clear protocols for data access and sharing[114,115]. Additionally, adopting federated learning approaches can enhance privacy by allowing model training across decentralized data sources without transferring sensitive patient information[116]. Such methodologies enable the development of accurate and personalized digital twins while minimizing privacy risks. Acknowledging and proactively addressing these privacy concerns is essential for the successful implementation of digital twins in precision health while maintaining patient trust and ensuring compliance with ethical and legal standards.

Conclusion

Digital twins represent a transformative approach in precision medicine, offering the potential to revolutionize health delivery by optimally tailoring interventions to improve a patient's current and future health states. However, the successful integration of digital twins into clinical workflows hinges on the rigorous application of the VVUQ process. By systematically addressing the challenges of model accuracy, computational efficiency, and the quantification and interpretation of the involved uncertainties, VVUQ will serve as the foundational framework that ensures safety, efficacy, and trustworthiness of digital twins.

To realize the full potential of digital twins in precision medicine, it is imperative that all stakeholders take concerted action. Researchers and developers need not only to incorporate robust VVUQ processes into the design and implementation of digital twins but also to invest in enhancing the interpretability of digital twin predictions and associated uncertainties for clinicians and patients. Clinicians, who guide the considerations around high-stakes decisions of their patients on a daily basis amid many uncertainties, possess valuable experience and expertise that are essential in setting expectations and standards for establishing confidence in digital twin predictions. The involvement of regulatory bodies is crucial to establish clear guidelines and standards for the VVUQ processes in digital twins, similar to those existing for clinical devices and trials. These guidelines must account for the unique characteristics of digital twins, including their dynamic nature that demands regulatory oversight to ensure safety and efficacy throughout their entire lifespan, not just at their initial deployment.

In this work, we examined the applications and implications of VVUQ processes for digital twins of precision medicine through specific examples in cardiology and oncology. We note that the scope of digital twins for precision medicine goes beyond these two key clinical application areas, where for instance, digital twins for psychiatry[117], orthopedics[118], pharmacy[119], or behavioral sciences[120], the findings and assessments may require adjustment. Additionally, the role for digital twins - as clinical decision rules versus clinical decision support - within the clinical workflow may vary across different focus areas in precision medicine. This variation requires a comprehensive survey, analyses and further research into the regulatory landscape, involving all stakeholders, including the FDA. Such efforts should also consider the digital twin's inputs, outputs, and the intended use. The integration of the VVUQ processes for ensuring accuracy and integrity of digital twin predictions should also maintain data security, protect individual privacy and address ethical demands at all stages.

By embracing these collective efforts, we can accelerate the adoption of digital twins in precision medicine and establish patient-specific health delivery as the norm rather than the exception.

Data availability

No datasets were generated or analysed during the current study.

Received: 21 October 2024; Accepted: 13 January 2025;
Published online: 17 January 2025

References

1. EDITH: European Virtual Human Twin. *Edith* https://www.edith-csa.eu/edith/ (2022).
2. National Academies of Sciences, E. & M. et al. *Foundational Research Gaps and Future Directions for Digital Twins*. (National Academies Press, Washington, D.C.,). https://doi.org/10.17226/26894 (2024).
3. Sel, K. et al. Building Digital Twins for Cardiovascular Health: From Principles to Clinical Impact. *J. Am. Heart Assoc.* **13**, 19 (2024).
4. Niederer, S. A., Campbell, K. S. & Campbell, S. G. A short history of the development of mathematical models of cardiac mechanics. *J. Mol. Cell Cardiol.* **127**, 11–19 (2019).
5. Liu, X.-Y. et al. Assessment of the effectiveness of BOPPPS-based hybrid teaching model in physiology education. *BMC Med Educ.* **22**, 217 (2022).
6. Ditterich, J. Stochastic models of decisions about motion direction: Behavior and physiology. *Neural Netw.* **19**, 981–1012 (2006).
7. Thorne, B. C., Hayenga, H. N., Humphrey, J. D. & Peirce, S. M. Toward a Multi-Scale Computational Model of Arterial Adaptation in Hypertension: Verification of a Multi-Cell Agent Based Model. *Front. Physiol.* **2**, 20 (2011).
8. Ricken, T. et al. Modeling function–perfusion behavior in liver lobules including tissue, blood, glucose, lactate and glycogen by use of a coupled two-scale PDE–ODE approach. *Biomech. Model Mechanobiol.* **14**, 515–536 (2015).
9. Skov, M. W. et al. Risk Prediction of Atrial Fibrillation Based on Electrocardiographic Interatrial Block. *J. Am. Heart Assoc.* **7**, 11 (2018).
10. Almutairi, H. M., Boubertakh, R., Miquel, M. E. & Petersen, S. E. Myocardial deformation assessment using cardiovascular magnetic resonance-feature tracking technique. *Br. J. Radio.* **90**, 20170072 (2017).
11. Krzyszczyk, P. et al. The growing role of precision and personalized medicine for cancer treatment. *Technol. (Singap. World Sci.)* **6**, 79–100 (2018).
12. Hinchey, M. G. & Bowen, J. P. *Applications of Formal Methods*. vol. 1 (Prentice Hall New York, 1995).

13. Rushby, J. Formal Methods and their Role in the Certification of Critical Systems. in *Safety and Reliability of Software Based Systems* 1–42 (Springer London, London,). https://doi.org/10.1007/978-1-4471-0921-1_1 (1997).
14. Oberkampf, W. L. & Trucano, T. G. Verification and validation in computational fluid dynamics. *Prog. Aerosp. Sci.* **38**, 209–272 (2002).
15. Oberkampf, W. L., Trucano, T. G. & Hirsch, C. Verification, validation, and predictive capability in computational engineering and physics. *Appl Mech. Rev.* **57**, 345–384 (2004).
16. Roy, C. J. Verification. in *Encyclopedia of Applied and Computational Mathematics* 1530–1537 (Springer Berlin Heidelberg, Berlin, Heidelberg,). https://doi.org/10.1007/978-3-540-70529-1_311 (2015).
17. Corrado, C. et al. Quantifying atrial anatomy uncertainty from clinical data and its impact on electro-physiology simulation predictions. *Med Image Anal.* **61**, 101626 (2020).
18. Brittain, H. K., Scott, R. & Thomas, E. The rise of the genome and personalised medicine. *Clin. Med (Lond.)* **17**, 545–551 (2017).
19. Gardiner, L.-J. et al. Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. *PLoS One* **17**, e0263248 (2022).
20. McMillan, C. T. MRI biomarkers — a precision medicine tool in neurology? *Nat. Rev. Neurol.* **12**, 323–324 (2016).
21. Hulsen, T. et al. From Big Data to Precision Medicine. *Front. Med. (Lausanne)* **6**, 34 (2019).
22. Sel, K. et al. Continuous cuffless blood pressure monitoring with a wearable ring bioimpedance device. *NPJ Digit Med* **6**, 59 (2023).
23. Kalantar-zadeh, K., Ha, N., Ou, J. Z. & Berean, K. J. Ingestible Sensors. *ACS Sens* **2**, 468–483 (2017).
24. Yacoub, M. H. & McLeod, C. The expanding role of implantable devices to monitor heart failure and pulmonary hypertension. *Nat. Rev. Cardiol.* **15**, 770–779 (2018).
25. Zhang, M., Adkins, M. & Wang, Z. Recent Progress on Semiconductor-Interface Facing Clinical Biosensing. *Sensors* **21**, 3467 (2021).
26. Sel, K., Mohammadi, A., Pettigrew, R. I. & Jafari, R. Physics-informed neural networks for modeling physiological time series for cuffless blood pressure estimation. *NPJ Digit Med* **6**, 110 (2023).
27. Lal, R. et al. Non Invasive Blood Flow Features Estimation in Cerebral Arteries from Uncertain Medical Data. *Ann. Biomed. Eng.* **45**, 2574–2591 (2017).
28. Gaidzik, F. et al. Hemodynamic Data Assimilation in a Subject-specific Circle of Willis Geometry. *Clin. Neuroradiol.* **31**, 643–651 (2021).
29. Chaudhuri, A. et al. Predictive digital twin for optimizing patient-specific radiotherapy regimens under uncertainty in high-grade gliomas. *Front. Artif. Intell.* **6**, 1222612 (2023).
30. Coveney, S. et al. Calibrating cardiac electrophysiology models using latent Gaussian processes on atrial manifolds. *Sci. Rep.* **12**, 16572 (2022).
31. Liu, S.-H., Yang, Z.-K., Pan, K.-L., Zhu, X. & Chen, W. Estimation of Left Ventricular Ejection Fraction Using Cardiovascular Hemodynamic Parameters and Pulse Morphological Characteristics with Machine Learning Algorithms. *Nutrients* **14**, 19 (2022).
32. Gupta, H. V., Wagener, T. & Liu, Y. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrol. Process* **22**, 3802–3813 (2008).
33. Eloranta, S. & Boman, M. Predictive models for clinical decision making: Deep dives in practical machine learning. *J. Intern Med* **292**, 278–295 (2022).
34. Amirahmadi, A., Ohlsson, M. & Etminani, K. Deep learning prediction models based on EHR trajectories: A systematic review. *J. Biomed. Inf.* **144**, 104430 (2023).
35. Rosen, I. I., Lane, R. G., Morrill, S. M. & Belli, J. A. Treatment plan optimization using linear programming. *Med Phys.* **18**, 141–152 (1991).
36. Abdar, M. et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021).
37. Schroeder, G. et al. Visualising the digital twin using web services and augmented reality. in *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)* 522–527 (IEEE,). https://doi.org/10.1109/INDIN.2016.7819217 (2016).
38. Liu, Y. K., Ong, S. K. & Nee, A. Y. C. State-of-the-art survey on digital twin implementations. *Adv. Manuf.* **10**, 1–23 (2022).
39. Nonnemann, L. et al. Health@Hand A Visual Interface for eHealth Monitoring. in *2019 IEEE Symposium on Computers and Communications (ISCC)* 1093–1096 (IEEE,). https://doi.org/10.1109/ISCC47284.2019.8969647 (2019).
40. Patel, R. J. et al. Use of surgical augmented intelligence maps can reduce radiation and improve safety in the endovascular treatment of complex aortic aneurysms. *J. Vasc. Surg.* **77**, 982–990.e2 (2023).
41. Liu, Y. et al. A Novel Cloud-Based Framework for the Elderly Healthcare Services Using Digital Twin. *IEEE Access* **7**, 49088–49101 (2019).
42. Shoukat, M. U. et al. Smart home for enhanced healthcare: exploring human machine interface oriented digital twin model. *Multimed. Tools Appl* **83**, 31297–31315 (2023).
43. Musen, M. A., Middleton, B. & Greenes, R. A. Clinical Decision-Support Systems. in *Biomedical Informatics* 795–840 (Springer International Publishing, Cham,). https://doi.org/10.1007/978-3-030-58721-5_24 (2021).
44. Chan, T. M. et al. Making Decisions in the Era of the Clinical Decision Rule: How Emergency Physicians Use Clinical Decision Rules. *Academic Med.* **95**, 1230–1237 (2020).
45. Helou, M. A., DiazGranados, D., Ryan, M. S. & Cyrus, J. W. Uncertainty in Decision Making in Medicine: A Scoping Review and Thematic Analysis of Conceptual Models. *Academic Med.* **95**, 157–165 (2020).
46. Heerink, J. S., Oudega, R., Hopstaken, R., Koffijberg, H. & Kusters, R. Clinical decision rules in primary care: necessary investments for sustainable healthcare. *Prim. Health Care Res Dev.* **24**, e34 (2023).
47. Valente, F. et al. Interpretability, personalization and reliability of a machine learning based clinical decision support system. *Data Min. Knowl. Discov.* **36**, 1140–1173 (2022).
48. Campbell-Washburn, A. E. et al. Opportunities in Interventional and Diagnostic Imaging by Using High-Performance Low-Field-Strength MRI. *Radiology* **293**, 384–393 (2019).
49. Milosevic, M. et al. Application of In Silico Trials for the Investigation of Drug Effects on Cardiomyopathy-Diseased Heart Cycle Properties. *Appl. Sci.* **13**, 11780 (2023).
50. Venkatesh, K. P., Raza, M. M. & Kvedar, J. C. Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation. *NPJ Digit Med* **5**, 150 (2022).
51. Shen, M., Chen, S. & Ding, X. The effectiveness of digital twins in promoting precision health across the entire population: a systematic review. *NPJ Digit Med* **7**, 145 (2024).
52. Katsoulakis, E. et al. Digital twins for health: a scoping review. *NPJ Digit Med* **7**, 77 (2024).
53. Sun, T., He, X. & Li, Z. Digital twin in healthcare: Recent updates and challenges. *Digit Health* **9**, 20552076221149652 (2023).
54. Thangaraj, P. M., Benson, S. H., Oikonomou, E. K., Asselbergs, F. W. & Khera, R. Cardiovascular care with digital twin technology in the era of generative artificial intelligence. *Eur Heart J* https://doi.org/10.1093/eurheartj/ehae619 (2024).
55. Lillie, E. O. et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per Med* **8**, 161–173 (2011).

56. Nye, L. Digital Twins for Patient Care via Knowledge Graphs and Closed-Form Continuous-Time Liquid Neural Networks. (2023).
57. Celebi, R. et al. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC Bioinforma.* **20**, 726 (2019).
58. Hosseini, S. et al. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Sci. Data* **9**, 255 (2022).
59. Spitzer, M., Dattner, I. & Zilcha-Mano, S. Digital twins and the future of precision mental health. *Front. Psychiatry* **14**, 1082598 (2023).
60. Ferdousi, R., Laamarti, F., Hossain, M. A., Yang, C. & El Saddik, A. Digital twins for well-being: an overview. *Digital Twin* **1**, 7 (2021).
61. Abilkaiyrkyzy, A., Laamarti, F., Hamdi, M. & El Saddik, A. Dialogue System for Early Mental Illness Detection: Toward a Digital Twin Solution. *IEEE Access* **12**, 2007–2024 (2024).
62. Hamilton, J. D. & Bickman, L. A Measurement Feedback System (MFS) Is Necessary to Improve Mental Health Outcomes. *J. Am. Acad. Child Adolesc. Psychiatry* **47**, 1114–1119 (2008).
63. Kapteyn, M. G., Pretorius, J. V. R. & Willcox, K. E. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nat. Comput Sci.* **1**, 337–347 (2021).
64. Davidson, K. W., Silverstein, M., Cheung, K., Paluch, R. A. & Epstein, L. H. Experimental Designs to Optimize Treatments for Individuals. *JAMA Pediatr.* **175**, 404 (2021).
65. Yankeelov, T. E. et al. Designing clinical trials for patients who are not average. *iScience* **27**, 108589 (2024).
66. Liu, R. et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).
67. Birnbaum, B. et al. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *CoRR* (2020).
68. Cai, Q., Zheng, K., Jagadish, H. V., Ooi, B. C. & Yip, J. CohortNet: Empowering Cohort Discovery for Interpretable Healthcare Analytics. (2024).
69. Hegselmann, S. et al. A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models. (2024).
70. Wang, Y. et al. TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model. (2024).
71. Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J. & Silva, R. Causal Machine Learning: A Survey and Open Problems. (2022).
72. Chen, J. W. et al. Best Practice Guidelines for Propensity Score Methods in Medical Research: Consideration on Theory, Implementation, and Reporting. A Review. *Arthrosc.: J. Arthroscopic Relat. Surg.* **38**, 632–642 (2022).
73. Vidovszky, A. A. et al. Increasing acceptance of AI-generated digital twins through clinical trial applications. *Clin. Transl. Sci.* **17**, e13897 (2024).
74. Mertens, J. & Denil, J. Reusing model validation methods for the continuous validation of digital twins of cyber-physical systems. *Softw Syst Model* https://doi.org/10.1007/s10270-024-01225-6 (2024).
75. Muñoz, P., Troya, J. & Vallecillo, A. Towards Measuring Digital Twins Fidelity at Runtime. in *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems* 507–512 (ACM, New York, NY, USA,). https://doi.org/10.1145/3652620.3688267 (2024).
76. Perisic, A. & Perisic, B. Digital Twins Verification and Validation Approach through the Quintuple Helix Conceptual Framework. *Electron. (Basel)* **13**, 3303 (2024).
77. Ogilvie, L. A., Kovachev, A., Wierling, C., Lange, B. M. H. & Lehrach, H. Models of Models: A Translational Route for Cancer Treatment and Drug Development. *Front. Oncol.* **7**, 219 (2017).
78. Baker, N. et al. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*. OSTI.GOV, 1478744 https://doi.org/10.2172/1478744 (2019).
79. Ahlquist, K. D., Sugden, L. A. & Ramachandran, S. Enabling interpretable machine learning for biological data with reliability scores. *PLoS Comput Biol.* **19**, e1011175 (2023).
80. Lee, C. et al. The UF family of reference hybrid phantoms for computational radiation dosimetry. *Phys. Med Biol.* **55**, 339–363 (2010).
81. Geyer, A. M., O'Reilly, S., Lee, C., Long, D. J. & Bolch, W. E. The UF/NCI family of hybrid computational phantoms representing the current US population of male and female children, adolescents, and adults—application to CT dosimetry. *Phys. Med Biol.* **59**, 5225–5242 (2014).
82. Salvador, M. et al. Whole-heart electromechanical simulations using Latent Neural Ordinary Differential Equations. *NPJ Digit Med* **7**, 90 (2024).
83. Molnar, C., Casalicchio, G. & Bischl, B. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. in *Joint European conference on machine learning and knowledge discovery in databases* 417–431 (Springer,). https://doi.org/10.1007/978-3-030-65965-3_28 (2020).
84. Fleeter, C. M., Geraci, G., Schiavazzi, D. E., Kahn, A. M. & Marsden, A. L. Multilevel and multifidelity uncertainty quantification for cardiovascular hemodynamics. *Comput Methods Appl Mech. Eng.* **365**, 113030 (2020).
85. Peherstorfer, B., Willcox, K. & Gunzburger, M. Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization. *SIAM Rev.* **60**, 550–591 (2018).
86. Loewe, A., Wülfers, E. M. & Seemann, G. Cardiac ischemia—insights from computational models. *Herzschrittmachertherapie + Elektrophysiologie* **29**, 48–56 (2018).
87. Kompa, B., Snoek, J. & Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* **4**, 4 (2021).
88. Phipps, J. et al. Early adverse physiological event detection using commercial wearables: challenges and opportunities. *NPJ Digit Med* **7**, 136 (2024).
89. Soleimany, A. P. et al. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).
90. MacDonald, S. et al. Generalising uncertainty improves accuracy and safety of deep learning analytics applied to oncology. *Sci. Rep.* **13**, 7395 (2023).
91. U. S. Food and Drug Administration. Clinical decision support software: guidance for industry and Food and Drug Administration staff. *FDA* https://downloads.regulations.gov/FDA-2017-D-6569-0114/attachment_1.pdf (2022).
92. Sutton, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* **3**, 17 (2020).
93. Shortliffe, E. H. & Sepúlveda, M. J. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* **320**, 2199 (2018).
94. U.S. Food and Drug Administration. Assessing the credibility of computational modeling and simulation in medical device submissions: guidance for industry and Food and Drug Administration staff. *FDA* https://downloads.regulations.gov/FDA-2021-D-0980-0001/attachment_1.pdf (2022).
95. U.S. Food and Drug Administration. Deciding when to submit a 510 (k) for a software change to an existing device: guidance for industry and Food and Drug Administration staff. *FDA* https://downloads.regulations.gov/FDA-2016-D-2021-0042/attachment_1.pdf (2017).
96. Morrison, T. M., Pathmanathan, P., Adwan, M. & Margerrison, E. Advancing Regulatory Science With Computational Modeling for Medical Devices at the FDA's Office of Science and Engineering Laboratories. *Front. Med. (Lausanne)* **5**, 241 (2018).

97. Taylor, C. A. et al. Patient-specific modeling of blood flow in the coronary arteries. *Comput Methods Appl Mech. Eng.* **417**, 116414 (2023).
98. U.S. Food and Drug Administration. HeartFlow Analysis (K213857): 510(k) premarket notification. *FDA* https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K213857 (2022).
99. Kołtowski, Ł. et al. Quantitative flow ratio derived from diagnostic coronary angiography in assessment of patients with intermediate coronary stenosis: a wire-free fractional flow reserve study. *Clin. Res. Cardiol.* **107**, 858–867 (2018).
100. Witberg, G. et al. Diagnostic Performance of Angiogram-Derived Fractional Flow Reserve. *JACC Cardiovasc Inter.* **13**, 488–497 (2020).
101. Levine, S. et al. Dassault Systèmes' Living Heart Project. in *Modelling Congenital Heart Disease* 245–259 (Springer International Publishing, Cham,). https://doi.org/10.1007/978-3-030-88892-3_25 (2022).
102. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
103. Kim, H.-E. et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* **2**, e138–e148 (2020).
104. Li, S., Zhao, R. & Zou, H. Artificial intelligence for diabetic retinopathy. *Chin. Med J. (Engl.)* **135**, 253–260 (2022).
105. Undru, T. R. et al. Integrating Artificial Intelligence for Clinical and Laboratory Diagnosis - a Review. *Maedica (Bucur.)* **17**, 420–426 (2022).
106. Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 1906–1919 (Association for Computational Linguistics, Stroudsburg, PA, USA,). https://doi.org/10.18653/v1/2020.acl-main.173. (2020).
107. Alowais, S. A. et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ.* **23**, 689 (2023).
108. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* **1**, e271–e297 (2019).
109. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* m689 https://doi.org/10.1136/bmj.m689 (2020).
110. Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. (2019).
111. Scott, I. A., van der Vegt, A., Lane, P., McPhail, S. & Magrabi, F. Achieving large-scale clinician adoption of AI-enabled decision support. *BMJ Health Care Inf.* **31**, e100971 (2024).
112. Afrose, S., Song, W., Nemeroff, C. B., Lu, C. & Yao, D. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Commun. Med.* **2**, 111 (2022).
113. Wentzel, A. et al. DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer. *IEEE Trans Vis Comput Graph* (2024).
114. Amofa, S. et al. Blockchain-secure patient Digital Twin in healthcare using smart contracts. *PLoS One* **19**, e0286120 (2024).
115. Alcaraz, C. & Lopez, J. Digital Twin: A Comprehensive Survey of Security Threats. *IEEE Commun. Surv. Tutor.* **24**, 1475–1503 (2022).
116. Gupta, D., Kayode, O., Bhatt, S., Gupta, M. & Tosun, A. S. Hierarchical Federated Learning based Anomaly Detection using Digital Twins for Smart Healthcare. (2021).
117. Rainey, S. Datafied Brains and Digital Twins: Lessons From Industry. *Caution Psychiatry Philos., Psychiatry, Psychol.* **29**, 29–42 (2022).
118. Aubert, K. et al. Development of Digital Twins to Optimize Trauma Surgery and Postoperative Management. A Case Study Focusing on Tibial Plateau Fracture. *Front Bioeng Biotechnol* **9**, (2021).
119. Subramanian, K. Digital Twin for Drug Discovery and Development—The Virtual Liver. *J. Indian Inst. Sci.* **100**, 653–662 (2020).
120. Schwartz, S. M., Wildenhaus, K., Bucher, A. & Byrd, B. Digital Twins and the Emerging Science of Self: Implications for Digital Health Experience Design and "Small" Data. *Front. Comput. Sci.* **2**, 31 (2020).

Acknowledgements

K.S., D.O., and R.J. acknowledge support from the National Institutes of Health (NIH) grants 7R01EB034821 and 7R01HL151240. A.C. and K.W. acknowledge support from the National Science Foundation (NSF) FDT-Biotech award 2436499 and the Defense Advanced Research Projects Agency (DARPA) Automating Scientific Knowledge Extraction and Modeling (ASKEM) program under award number DE-AC05-76RL01830. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

Author contributions

K.S., A.H.D., A.C., D.O., A.B., D.P., K.W., C.C., and R.J. contributed to the conception of the ideas presented in this framework. K.S., A.H.D., A.C., D.O., and R.J. performed the systematic survey of the literature on VVUQ for digital twins. The manuscript was written with the contributions of all authors. All authors have approved the final version of the manuscript.

Competing interests

R.J. serves as an Associate Editor for npj Digital Medicine and is a cofounder of SpectroBeat, LLC. A.B. serves as the Chief Technology Officer of Texas Instruments Incorporated. Other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Roozbeh Jafari.