# Leveraging Electronic Health Records and Mechanistic Models for Predictive Digital Twins in Healthcare

## 1. Introduction: The Emergence of EHR-Based Digital Twins in Healthcare

The concept of the Digital Twin (DT), a virtual representation dynamically linked to a physical counterpart, has migrated from its origins in engineering and industry 1 to become a focal point of innovation within the healthcare sector.1 This transition holds immense promise for transforming multiple facets of medicine, including patient care delivery, the optimization of clinical operations, and the acceleration of medical research.2 Healthcare DTs, particularly those constructed using patient data primarily sourced from Electronic Health Records (EHRs) 5, offer the potential to create highly personalized, predictive models of individual health trajectories.

This report provides a deep research background on the specific application of EHR-based DTs for predictive purposes, with a particular emphasis on the integration of mechanistic models. Mechanistic models, which aim to represent underlying biological or pathophysiological processes 2, are increasingly seen as crucial for enhancing the predictive accuracy, interpretability, and clinical actionability of DTs beyond what purely data-driven approaches can offer. The ultimate goal is often framed within the paradigm of precision medicine, tailoring prevention and treatment strategies to the unique characteristics of each individual.2

The rapid advancement of healthcare DTs is not occurring in isolation. It represents a significant convergence of several powerful trends that are reshaping the biomedical landscape. The widespread adoption of EHR systems provides an unprecedented, longitudinal data foundation.5 Concurrently, breakthroughs in artificial intelligence (AI) and machine learning (ML) offer sophisticated tools for analyzing complex health data.2 The proliferation of the Internet of Things (IoT), including wearable sensors and connected medical devices, generates streams of real-time physiological data.2 Furthermore, increases in computational power make complex simulations feasible 1, while the overarching drive towards precision medicine creates a strong clinical pull for personalized predictive tools.2 The DT concept 1 serves as an integrating framework, bringing these disparate elements together into a cohesive structure aimed at understanding and managing individual health. Consequently, progress in healthcare DTs is intrinsically linked to advancements in these underlying areas; challenges such as data interoperability, AI transparency, or sensor accuracy inevitably act as constraints on DT development and deployment.

This report will systematically explore the landscape of EHR-based predictive DTs incorporating mechanistic understanding. It begins by defining the healthcare DT and distinguishing it from related concepts. Subsequent sections delve into the integration of mechanistic models, specific predictive applications and case studies (with a focus on psychiatry), the critical data requirements and associated challenges inherent in using EHRs, the computational methodologies employed, strategies for validation and assessing clinical readiness, and the crucial ethical, legal, and social implications (ELSI). Finally, the report identifies current limitations and outlines promising future directions for research and development in this rapidly evolving field.

## 2. Defining the Healthcare Digital Twin

### 2.1 Origins and Evolution

The term "digital twin" traces its conceptual origins to industry and engineering. Early ideas emerged at the University of Michigan in 2002, outlining real and virtual spaces with bidirectional information flow.1 The term itself gained prominence with a 2012 NASA report defining it as an "integrated multiphysics, multiscale, probabilistic simulation of an as-built vehicle or system" linked to its physical counterpart via sensor updates to mirror its life.1 This industrial concept, driven by advances in cyber-physical systems (where computational elements control physical entities) and high-fidelity simulation 1, has since expanded into numerous fields, including healthcare.1

In the medical domain, the DT concept has evolved from representing static replicas to encompassing dynamic, adaptive, and even "intelligent" virtual counterparts capable of learning and reasoning.4 A specialized variant, the Digital Human Twin (DHT), specifically refers to virtual replicas focused on human physiology for healthcare applications.6

### 2.2 Core Components

Despite variations in application, a consensus has emerged around the core components defining a healthcare DT 2:

1. **The Real Entity:** This is the physical object or system being mirrored. In healthcare, this is typically the patient, but can also be a specific organ (e.g., heart, tumor), tissue, cell population, or even a broader system like a hospital workflow or environment.2 The term "Real Entity" is sometimes preferred over "Physical Entity" as the target may encompass processes or workflows that are not strictly physical objects.3
2. **The Virtual Representation (Replica):** This is the digital counterpart, a computational model designed to accurately represent the real entity. It can capture various dimensions, including geometry, physics, behavior, and potentially complex multi-scale and multi-physics interactions.1 This goes beyond static visualizations like 3D models to encompass dynamic function.3
3. **The Data Connection (Bidirectional Flow):** This is the critical communication channel linking the real entity and its virtual representation.1 Data (e.g., from EHRs, sensors, imaging) flows from the real entity to continuously update the virtual model. Information, predictions, or potentially even control signals can flow back from the virtual twin to inform actions or decisions related to the real entity. This live, dynamic connection is considered a primary differentiator from traditional modeling.3

It is important to recognize that the nature of the "bidirectional" flow exists on a spectrum.3 Some conceptualizations envision a fully integrated system where the virtual twin might directly influence or control aspects of the physical counterpart.3 A more common interpretation, particularly in current healthcare applications, involves a continuous or near real-time flow of data *to* the virtual twin to update its state, while the flow *back* consists of information, predictions, or simulation results that inform clinician or patient decisions regarding the real entity.3 This distinction carries significant weight, influencing the potential applications (ranging from enhanced monitoring and decision support to, perhaps in the future, more automated interventions) and the associated regulatory and ethical considerations. Clarifying the specific nature and agency of this feedback loop is essential when evaluating any DT system.

### 2.3 Key Characteristics

Healthcare DTs are typically characterized by several key features:

* **Dynamic and Continuously Updated:** Unlike static models, DTs are designed to evolve over time, reflecting changes in the real entity based on incoming data streams.1 This continuous updating process suggests that a DT is perhaps better understood not just as a static object or replica, but as a dynamic *process* of mirroring, predicting, comparing predictions to reality, and refining the model in lockstep with the patient's health journey.1 This perspective underscores the need for robust infrastructure capable of handling continuous data ingestion, processing, and model recalibration, and implies that the accuracy or fidelity of the twin is inherently time-dependent.
* **Personalized/Individualized:** DTs are tailored to a specific individual, integrating their unique data from sources like EHRs, genomic sequencing, wearable sensors, imaging, and patient-reported outcomes.2
* **Predictive:** A primary function is to generate forecasts about future health states, such as disease onset or progression, response to potential treatments, or the likelihood of adverse events.2
* **The "Five I's":** Deng proposed five characteristics of an effective DT: **Individualized** (patient-specific), **Interconnected** (real-time link between physical and virtual), **Interactive** (allows simulation and 'what-if' scenarios), **Informative** (provides actionable insights), and **Impactful** (enhances quality of life or health outcomes).40

### 2.4 Distinction from Traditional Models

It is crucial to differentiate DTs from related but distinct concepts like simulation and standalone predictive models:

* **Simulation vs. Digital Twin:** Traditional simulation often models hypothetical scenarios – what *could* happen under certain conditions. It may use predefined datasets and is often static, meaning the model doesn't change unless manually altered by a designer.44 While DTs frequently *incorporate* simulation techniques 1, they are fundamentally defined by the dynamic, typically real-time, bidirectional data connection to a *specific* real-world entity.3 A DT aims to replicate what *is* happening to its physical counterpart, continuously updating based on real data feed, making it an active and evolving representation.41 The analogy of a static city map versus a dynamic GPS system effectively captures this difference.48
* **Predictive Models (Standalone) vs. Digital Twin:** Standard predictive models in healthcare often use historical (often static) data to predict a specific future outcome (e.g., risk of readmission). While DTs certainly employ predictive modeling as a core function 12, they are distinguished by their holistic approach. A DT aims to create a comprehensive, dynamic virtual representation of the patient (or system) that mirrors its state and behavior over time, enabled by the continuous data feed.12 The prediction is generated *from* this dynamic representation, which is constantly being updated. An AI or ML model applied to EHR data for patient stratification, for instance, might not qualify as a DT unless it is integrated within this framework of a persistent virtual replica with a live data connection.41

## 3. Integrating Mechanistic Understanding with EHR Data

### 3.1 The Rationale for Mechanistic Integration

While purely data-driven approaches, such as standard machine learning algorithms applied to EHR data, have shown success in identifying correlations and making predictions based on past patterns, they often face limitations. These models can struggle to predict responses to novel interventions or situations not well-represented in the training data, and they typically provide limited insight into the underlying biological reasons for a prediction ('black box' problem).25 This lack of mechanistic grounding can hinder clinical trust and the ability to generate truly actionable information for personalized treatment decisions.24

Integrating mechanistic models offers a path to overcome these limitations. These models aim to explicitly represent the underlying biological, physiological, or pathophysiological processes driving health and disease.2 By incorporating knowledge of how biological systems work, mechanistic DTs strive to provide not just predictions, but also explanations, enabling a deeper understanding of individual patient dynamics. This mechanistic foundation is particularly valuable for simulating the effects of interventions (e.g., drugs, therapies) by modeling how they perturb specific biological pathways or processes.25 This capability transforms the DT from a purely correlative tool into a platform for conducting *in silico* experiments and testing hypotheses about an individual patient's likely response to different therapeutic strategies.25 Such a framework allows the DT to explore counterfactuals ("What would happen if this treatment were administered?") based on biological principles, rather than solely relying on patterns observed in historical EHR data. This potential for extrapolation beyond observed data and enhanced interpretability is key to building clinician trust and enabling more precise, individualized interventions.

### 3.2 Types of Mechanistic Models in Healthcare DTs

Various types of mechanistic models are being explored for integration into healthcare DTs:

* **Physiological Pathway Models:** These models often use systems of mathematical equations, such as Ordinary Differential Equations (ODEs) or Partial Differential Equations (PDEs), to describe the dynamics of biochemical reactions, signaling cascades, gene regulatory networks, or metabolic pathways.14 They aim to capture the interactions between molecular components over time.14
* **Disease Progression Models:** These focus on simulating the natural history of a disease or its response to treatment over longer timescales, capturing key stages and transitions.6
* **Multi-Scale Models:** Recognizing that biological processes span multiple levels of organization, these models attempt to bridge scales from the molecular (e.g., drug-target interactions) to the cellular (e.g., cell signaling, fate decisions), tissue (e.g., growth, remodeling), organ (e.g., heart electrophysiology, tumor microenvironment), and even the whole-body or population level.1 This is critical because many interventions target mechanisms at one scale (e.g., intracellular) but exert their clinically relevant effects at higher scales (e.g., tissue or organ).25
* **Specific Modeling Paradigms:** Beyond ODEs/PDEs, other paradigms include Agent-Based Models (ABMs), which simulate the behavior and interactions of individual autonomous agents (like cells) 25; Boolean Networks, representing logical relationships in regulatory networks 25; and Knowledge Graphs that integrate mechanistic understanding with data-derived relationships.25

### 3.3 Synergies and Integration with EHR Data

EHR data, despite its limitations, provides the crucial real-world context and longitudinal information needed to ground, personalize, calibrate, and validate these mechanistic models for individual patients.2 Integration can be achieved through several approaches:

* **Hybrid Modeling:** This involves combining data-driven components (often ML/statistical models trained on EHR data) with mechanistic modules.2 For example, ML might be used to extract complex phenotypes or risk scores from EHRs, which then serve as inputs or constraints for a mechanistic simulation of disease progression or treatment response.
* **Data Assimilation:** Techniques used to continuously or periodically update the state variables or parameters of a running mechanistic model using new data points extracted from the EHR (e.g., recent lab results, vital signs).25
* **Model Personalization ("Twinning"):** Patient-specific data derived from the EHR (e.g., demographics, baseline diagnoses, comorbidities, key lab values, findings extracted from imaging or pathology reports via NLP) are used to set the initial conditions, boundary conditions, or specific parameter values of a generic mechanistic model, thereby creating a personalized instance.25

### 3.4 Challenges in Mechanistic Integration

Despite the promise, integrating mechanistic models with EHR data for DTs faces significant hurdles:

* **Biological Complexity and Knowledge Gaps:** Human biology is incredibly complex, characterized by non-linearity, emergent properties, heterogeneity, and stochasticity.25 Unlike engineered systems, biological systems are products of evolution, and our understanding of many underlying mechanisms remains incomplete.25 This often makes it impossible to build fully specified, purely mechanistic models for many conditions, necessitating reliance on data-driven approaches or significant simplifying assumptions.25
* **The Data-Knowledge Chasm:** A fundamental tension exists between the desire for deep mechanistic insight, which requires detailed biological knowledge and often granular data, and the reality of EHR data, which is primarily collected for clinical care and billing, often resulting in noise, incompleteness, non-standardization, and a lack of specific biological measurements needed by detailed models.7 Building high-fidelity mechanistic DTs solely from standard EHR data is frequently infeasible. Practical approaches often involve hybrid models where EHR data informs certain aspects (e.g., overall patient state, comorbidities), while mechanistic components rely on existing biological knowledge, which may itself be incomplete. This highlights the need for methods that can gracefully integrate partial knowledge with imperfect data. Causal inference techniques 50, which attempt to uncover cause-effect relationships from observational data like EHRs, represent one potential avenue to bridge this gap, approximating mechanistic links without requiring full biological specification.
* **Model Complexity vs. Data Resolution:** There is a trade-off between the desired granularity of the mechanistic model and the resolution and quality of data available from EHRs.25 Highly detailed models may require data not routinely collected, while models simplified to match EHR data granularity might lose crucial mechanistic detail. Furthermore, even if data were available, highly granular predictions might not be clinically actionable with current therapeutic options.25
* **Computational Cost:** Multi-scale, complex mechanistic simulations can be computationally demanding, potentially hindering real-time updates or rapid exploration of multiple scenarios required for clinical decision support.9 Faster surrogate models, perhaps trained using ML, may be needed.14
* **Bridging Scales:** Developing models that accurately link processes across different biological scales (e.g., molecular interactions influencing organ function) remains a significant technical challenge.14

## 4. Predictive Applications and Clinical Case Studies

EHR-based digital twins, particularly those incorporating mechanistic insights, are being developed and explored for a range of predictive applications across various clinical domains. The primary goals often involve moving beyond population-level statistics to individualized forecasting:

* **Disease Onset/Risk Prediction:** Identifying individuals within a population who are at elevated risk of developing specific conditions in the future, enabling targeted screening or preventative interventions.2
* **Disease Progression Forecasting:** Simulating the likely future course of an existing disease for a specific patient, potentially under different management scenarios.6
* **Treatment Response Prediction:** Forecasting how an individual patient is likely to respond to a particular drug, therapy, or intervention, aiding in the selection of the most effective and least harmful treatment strategy.2
* **Adverse Event Prediction:** Identifying patients at high risk for experiencing specific complications, side effects from treatment, or other adverse outcomes.15

### 4.1 Focus Domain: Psychiatry (Substance Use, Anxiety/Depression)

Psychiatry presents both significant challenges and opportunities for EHR-based predictive modeling and DTs. Challenges include the inherent heterogeneity within diagnostic categories like Major Depressive Disorder (MDD) 55, the limitations of purely categorical diagnostic systems that fail to capture symptom overlap or dimensional aspects of psychopathology 55, and the fact that crucial information, such as substance use details or nuanced therapy outcomes, is often documented primarily in unstructured clinical notes rather than structured EHR fields.56 The Research Domain Criteria (RDoC) framework was introduced to encourage dimensional approaches linking symptoms to biology, but assessing these dimensions clinically remains difficult.55

Despite these challenges, EHR data, encompassing both structured elements (diagnoses, medications, demographics) and unstructured text, is being actively leveraged:

* **NLP is Essential:** Natural Language Processing (NLP) techniques are proving indispensable for extracting vital information from clinical notes.51 This includes identifying specific symptoms, assessing sentiment or emotional states, detecting mentions of substance use, recognizing suicidal ideation, and capturing nuanced outcomes of psychotherapy.53 Multiple studies demonstrate that incorporating NLP analysis of notes significantly improves predictive performance compared to using structured data alone.53 This underscores that for complex fields like psychiatry, relying solely on structured EHR data is likely insufficient, making robust NLP capabilities a prerequisite for effective EHR-based DTs or predictive models. The quality and consistency of clinical documentation thus become critically important.
* **Predictive Modeling Examples:**
  + **Depression/Anxiety Prediction:** A deep learning framework (DAP) successfully predicted post-discharge depression and anxiety in Type 2 Diabetes Mellitus patients using regional EHR data, employing unsupervised contrastive pre-training to handle heterogeneous data.51 Another study developed regression models using structured EHR data and free-text notes to predict a future diagnosis of depression up to 12 months in advance (AUC 0.70–0.80) and differentiate severe from minimal/mild baseline depression (AUC 0.72).52 This study also explored predicting differential response to medication versus psychotherapy, finding baseline severity to be the strongest predictor.52
  + **Substance Use Detection:** An automated system (ASUDS) integrated logic rules for structured EHR data with NLP (knowledge-based and deep learning) for clinical notes to detect substance use (alcohol, tobacco, marijuana, opiates) in pediatric patients. The integrated approach achieved high performance (AUC 0.96) and captured significantly more positive cases than structured data alone (94.0% vs 22.0%).57 NLP is also used more broadly to identify discussions related to drug use and addiction in various text sources.58
  + **Suicide Risk Prediction:** A large data-driven approach using ML and NLP on 2 years of historical EHR data (structured and unstructured notes) accurately predicted first-time suicide attempts across various prediction windows (AUC up to 0.932). The inclusion of NLP-derived features from notes significantly outperformed models using only structured data.53
  + **Clinical Decision Support (CDS) Implementation:** An AI-driven CDS system for postpartum depression (PPD) prediction, trained on EHR data and refined using a consortium dataset for generalizability and fairness, was deployed using Microsoft Azure and FHIR standards. Risk assessments were integrated into the clinician's workflow.59

### 4.2 Applications in Other Clinical Domains

While psychiatry highlights the importance of NLP, EHR-based DTs and predictive models are advancing in other areas as well:

* **Oncology:** Significant effort is focused on creating Cancer Patient Digital Twins (CPDTs) that integrate clinical, genomic/multi-omic, imaging, and potentially lifestyle data.26 These aim to predict cancer progression and individual response to therapies like chemotherapy.15 Frameworks like FarrSight®-Twin leverage clinical and molecular profiles for prediction.54 Proposed uses include performing virtual experiments to compare treatment strategies 26, optimizing treatment selection 54, and potentially serving as synthetic control arms in clinical trials.35
* **Cardiology:** DTs are used to model heart function, incorporating data from imaging (MRI, CT), ECG, EHRs, and wearables.3 Applications include predicting response to interventions (e.g., stent placement, cardiac resynchronization therapy), forecasting heart failure risk, simulating cardiac electrophysiology to understand and plan treatments for arrhythmias (e.g., atrial fibrillation, ventricular tachycardia), and improving risk stratification for events like sudden cardiac death.3
* **Other Areas:** Beyond patient-specific prediction, DT concepts are applied to optimize hospital operations (resource allocation, patient flow, staffing) 5, model infectious disease spread during pandemics 12, and improve processes in biomanufacturing and pharmaceuticals.12

### 4.3 Table: Summary of Predictive Applications using EHR-based Models/DTs

Table 4.1 provides a synthesized overview of representative predictive applications discussed in the source materials.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Clinical Domain** | **Prediction Goal** | **Key Methods/Data** | **Key Findings/Performance** | **Snippet Reference(s)** |
| Psychiatry (Depression/Anxiety) | Predict Post-Discharge D/A in T2DM Patients | Deep Learning (contrastive pre-training) on regional EHRs (DAP Model) | External Validation: AUC 0.75, PR-AUC 0.47. Outperformed baselines. | 51 |
| Psychiatry (Depression) | Predict Future Diagnosis (≤12 mo); Assess Severity | Regression on Structured EHR + Free-text Notes (NLP) | Diagnosis Prediction: AUC 0.70-0.80. Severity Differentiation: AUC 0.72. | 52 |
| Psychiatry (Substance Use) | Detect Substance Use Screening/Results (Pediatric) | Logic Rules (Structured EHR) + NLP/Deep Learning (Notes) | Integrated AUC 0.96. NLP vital (captured 94% positives vs 22% structured). | 57 |
| Psychiatry (Suicide Risk) | Predict First-Time Suicide Attempt | ML + NLP on Structured EHR + Notes | AUC 0.932 (Combined) vs 0.901 (Structured only). Robust across windows/subgroups. | 53 |
| Psychiatry (Postpartum Depression) | Predict PPD Risk for CDS | AI model on EHR data (refined with consortium data) | Deployed via Azure/FHIR into clinical workflow. (Performance metrics not detailed). | 59 |
| Oncology | Predict Chemotherapy Response; Simulate Trials | CPDT framework; FarrSight®-Twin (Clinical + Molecular Profiles); AI/Mechanistic | Concordance in simulated trials; Improved ORR/OS in enriched cohorts; Synthetic controls | 26 |
| Cardiology | Predict Treatment Response (Interventions); Risk Strat. | Mechanistic Models + EHR, Imaging (MRI/CT), ECG, Wearables | Personalized models for arrhythmia care (CRT, AF, VT), heart failure risk prediction. | 3 |

A recurring theme across these applications is the journey from prediction to intervention. While many studies demonstrate promising accuracy in predicting diagnoses, risks, or even treatment responses based on EHR data 51, there is often a gap in demonstrating that these predictions translate into effective clinical interventions that demonstrably improve patient outcomes. The PPD case study 59, for example, describes deployment into workflow, but the impact on PPD rates or management is not detailed in the available information. Similarly, frameworks for CPDTs envision decision support 26, but evidence of improved outcomes from prospective application is still emerging. This highlights a critical need for future research to move beyond model development and validation towards rigorous evaluation of the clinical utility and real-world impact of DT-informed healthcare. Prediction is a necessary first step, but demonstrating improved patient health is the ultimate goal.

## 5. Data Ecosystem: Requirements and Challenges for EHR-Based Digital Twins

Constructing robust and reliable EHR-based digital twins necessitates a rich and complex data ecosystem. However, leveraging this ecosystem presents significant practical challenges.

### 5.1 Data Requirements

The ideal DT integrates diverse data streams to create a holistic, multi-dimensional view of the patient:

* **Core EHR Data:** Both structured data (e.g., demographics, diagnoses (ICD codes), procedures (CPT codes), medications, lab results) and unstructured data (e.g., clinical notes, discharge summaries, imaging/pathology reports) are fundamental.5 NLP is crucial for unlocking unstructured data.53
* **Genomics and Multi-Omics:** Data such as genomics, proteomics, metabolomics provide deep biological insights relevant for personalization, particularly in areas like oncology and precision medicine.5
* **Medical Imaging:** Modalities like CT, MRI, PET provide anatomical and functional information, crucial for modeling specific organs or tissues.7
* **Wearables and IoT Devices:** Continuous physiological data (e.g., heart rate, activity levels, glucose monitoring) from wearables and remote sensors offer real-time insights into patient state and behavior outside clinical settings.2
* **Patient-Reported Data:** Information directly from patients about symptoms, quality of life, and experiences provides valuable subjective context.11
* **Environmental and Social Determinants of Health (SDoH):** Factors like socioeconomic status, environment, lifestyle, and access to resources significantly impact health and should ideally be incorporated.9 This data might come from EHR notes, patient surveys, or external area-level datasets.63

Beyond diversity, the data must meet stringent quality standards:

* **Quality:** Data needs to be accurate, complete, consistent, and timely.7 However, EHR data quality is notoriously variable, with historical reviews showing wide ranges in correctness and completeness.64 Ensuring data integrity is critical for reliable DT models.39
* **Standardization:** Use of standardized metadata, ontologies (e.g., SNOMED CT, LOINC), and controlled clinical terminologies is essential to avoid disorganized "data swamps" and enable meaningful integration.9
* **Volume:** While not always paramount ('appropriately sized' data is key 39), large datasets are often required, particularly for training complex data-driven components (e.g., deep learning models) and ensuring statistical power.2

### 5.2 Critical Evaluation of Challenges

Leveraging EHR data for DTs involves navigating numerous challenges, which can also be viewed as opportunities for methodological innovation:

* **Incompleteness and Missing Data:** EHRs are frequently incomplete, lacking data points expected or needed for specific analyses.9 Completeness was identified as the most commonly assessed data quality dimension in a review of EHR data quality studies.64 This necessitates methods for robust handling of missing information.
* **Heterogeneity:** Data originates from diverse sources, systems, and formats, encompassing structured fields, free text, images, signals, and -omics data, making integration complex.7
* **Bias:** EHR data can reflect and perpetuate existing biases in healthcare access, diagnosis, treatment, and documentation practices. Models trained on such data risk inheriting these biases, potentially leading to unfair or inequitable DT predictions and exacerbating health disparities.9 Addressing algorithmic bias is a critical ELSI concern.9
* **Privacy Constraints and Security:** Protecting highly sensitive patient health information is a paramount legal and ethical obligation.7 Ensuring compliance with regulations like HIPAA and GDPR, implementing robust security measures (encryption, access control), and developing privacy-preserving techniques are essential.7 Meaningful anonymization is increasingly difficult in the face of rich, multi-modal data.67 This creates a fundamental tension: the drive for deep personalization inherent in the DT concept requires integrating vast amounts of potentially identifying data, directly conflicting with privacy imperatives. The more personalized the twin, the greater the potential privacy risk. Consequently, robust privacy engineering (e.g., differential privacy, federated learning, secure computation) and strong governance are not optional but foundational requirements for ethical DT deployment.
* **Interoperability Issues:** The lack of universally adopted standards for data representation and exchange protocols significantly hinders the ability to seamlessly integrate data from different EHR systems, medical devices, and other sources.7 This limits scalability and data sharing across institutions. Efforts promoting standards like FHIR (Fast Healthcare Interoperability Resources) 59 are crucial.
* **Data Accessibility:** Simply gaining access to the necessary data, which may reside in siloed systems within or across healthcare organizations, can be a major logistical and administrative barrier.7
* **Cost and Effort:** The processes of acquiring, cleaning, standardizing, integrating, and managing the large, complex datasets required for DTs are resource-intensive, demanding significant time, expertise, and financial investment.39

These data challenges are not merely technical hurdles; they are often systemic, deeply embedded in clinical documentation practices, healthcare organizational structures, and the historical lack of enforced data standards.7 Technical solutions like advanced algorithms are necessary but insufficient. Addressing the data foundation for DTs requires a multi-faceted strategy involving technological innovation, policy mandates for standardization (like the FAIR - Findable, Accessible, Interoperable, Reusable - principles 26), adoption of common data models 68, potential redesign of clinical workflows to improve data capture, and strong institutional commitment.

### 5.3 Table: EHR Data Challenges and Mitigation Opportunities

Table 5.1 summarizes key data challenges and potential avenues for innovation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Challenge** | **Description** | **Example Snippets** | **Potential Mitigation Strategies/Opportunities** | **Example Snippets for Mitigations** |
| Incompleteness | Missing data points in EHRs due to documentation gaps or inconsistent collection. | 9 | Advanced imputation methods; ML models robust to missingness; NLP to extract data from notes; Synthetic data generation. | 12 |
| Heterogeneity | Data from diverse sources, formats (structured, unstructured, imaging, omics), and systems. | 7 | Data standardization (ontologies, terminologies); Common Data Models; Multi-modal ML/AI integration techniques; Data harmonization pipelines. | 9 |
| Bias | EHR data reflecting underlying biases in care access, delivery, or documentation, leading to biased models. | 9 | Bias detection algorithms; Fairness-aware ML; Careful cohort selection; Diverse data sourcing; Post-hoc bias correction; Transparency. | 9 |
| Privacy/Security | Risk of re-identification and unauthorized access to sensitive patient health information. | 7 | Privacy-preserving techniques (Federated Learning, Differential Privacy, Encryption); Robust access controls; Compliance frameworks (HIPAA/GDPR); Synthetic data; Privacy risk assessment. | 7 |
| Interoperability | Lack of standards hindering seamless data exchange and integration between systems. | 7 | Adoption of standards (FHIR, CDS Hooks); Development of interoperable platforms; Open APIs; Semantic interoperability efforts. | 7 |
| Accessibility | Difficulties in accessing required data across institutional or system silos. | 7 | Data sharing agreements; Research networks/consortia; Federated data access models; Patient-mediated data sharing. | 7 |
| Data Quality | Issues with accuracy, consistency, timeliness of EHR data. | 41 | Data quality assessment tools; Automated data cleaning pipelines; Improved clinical documentation practices; Feedback loops to clinicians. | 9 |

## 6. Computational Methodologies for Development and Utilization

Developing and utilizing EHR-based digital twins involves a diverse array of computational and analytical methodologies, often employed in combination to leverage their respective strengths.

### 6.1 Overview of Key Techniques

* **Machine Learning (ML):** ML serves as a cornerstone for extracting patterns, building predictive models, and analyzing complex relationships within large EHR datasets.4 This includes a wide range of algorithms, such as logistic regression, random forests, support vector machines, and gradient boosting machines (e.g., XGBoost).51
* **Deep Learning (DL):** As a subset of ML, DL techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformers, are increasingly applied, particularly for handling high-dimensional, heterogeneous, and sequential EHR data (like clinical notes or time-series measurements).22 Unsupervised contrastive pre-training on large EHR datasets can help models learn robust representations before fine-tuning on specific tasks.51 DL is also well-suited for integrating multi-modal data inputs.36
* **Causal Inference Methods:** Recognizing the limitations of purely correlational ML, causal inference techniques aim to estimate the causal effects of interventions (e.g., treatments) from observational EHR data, controlling for confounding factors.50 This is crucial for moving towards personalized treatment recommendations and understanding *why* certain outcomes occur.50 These methods provide tools for reasoning about the potential results of actions, bridging the gap between prediction and actionable clinical insight. While challenging to apply correctly due to reliance on untestable assumptions, they offer a pathway to extract more mechanistic understanding from observational data.
* **Statistical Modeling:** Traditional statistical methods, including regression analysis, survival analysis, and time-series modeling, remain essential for data analysis, hypothesis testing, model validation, and complementing ML approaches.2
* **Simulation Techniques:** Simulation is often embedded within the DT framework to explore dynamic processes, test 'what-if' scenarios, and predict future trajectories.1 This includes methods like agent-based modeling (ABM) 25, discrete-event simulation (often used for operational modeling 71), and simulations based on mechanistic models (e.g., ODE/PDE systems).14
* **Large Language Models (LLMs):** A newer frontier involves leveraging LLMs (e.g., GPT variants) for DT creation and prediction.22 Approaches like TWIN-GPT propose using the vast clinical knowledge embedded in LLMs to generate personalized patient simulations, potentially overcoming data gaps and inconsistencies in EHRs.46 LLMs can establish cross-dataset associations and perform zero-shot forecasting on variables not explicitly seen during training.22 This represents a potential paradigm shift, moving from models built solely on local EHR data towards leveraging large, pre-trained foundational models infused with external knowledge. However, this introduces challenges related to controlling outputs, ensuring factual accuracy (mitigating 'hallucination'), quantifying uncertainty, and the need for careful prompt engineering and fine-tuning.46
* **Generative Models:** Techniques like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are used primarily for generating synthetic EHR data, which can be valuable for augmenting datasets, testing models, or preserving privacy.12

### 6.2 Strengths and Weaknesses

Each methodology brings distinct advantages and disadvantages to the task of building EHR-based DTs:

* **ML/DL:** Excel at identifying complex, non-linear patterns in large EHR datasets without requiring pre-specified biological knowledge. However, they often function as 'black boxes,' making interpretation difficult, may struggle with causality, can be sensitive to data biases, and typically require substantial amounts of labeled training data.28
* **Causal Inference:** Offers a framework for estimating the effects of specific actions or interventions from observational data, providing more actionable insights than pure correlation. However, methods rely on strong, often untestable assumptions about the data generating process (e.g., no unmeasured confounding), and can be complex to implement and validate correctly.
* **Mechanistic Models:** Provide interpretability based on biological understanding, allow for extrapolation beyond observed data patterns, and facilitate *in silico* hypothesis testing about interventions. However, they are limited by gaps in biological knowledge, can be computationally expensive, and may require detailed data not always available in EHRs.14
* **Simulation:** Enables exploration of dynamic system behavior and complex interactions over time. However, building accurate simulation models can be challenging, and calibrating and validating them against sparse or noisy real-world EHR data is often difficult.
* **LLMs:** Can leverage vast pre-existing knowledge to potentially overcome data scarcity and generate highly personalized outputs. However, they face challenges with factual consistency (hallucination), controllability, transparency, robust uncertainty quantification, and require careful tuning and validation for specific medical tasks.46

### 6.3 Hybrid Approaches

Given the complementary strengths and weaknesses, hybrid approaches that combine different methodologies are often the most promising strategy for building effective DTs.2 For instance, ML can be used to process raw EHR data and extract key features or patient states, which then inform or parameterize a mechanistic model simulating core biological processes. Causal inference might refine understanding of treatment effects within the model. Simulation allows for exploring the dynamic evolution predicted by the hybrid model under different scenarios. This integration aims to create DTs that are both data-grounded and mechanistically informative.

## 7. Validation Strategies and Assessing Clinical Readiness

The translation of EHR-based digital twins from research concepts to clinically useful tools hinges on rigorous validation and a clear assessment of their readiness for real-world application. Given their potential to influence critical patient care decisions, the validation requirements for medical DTs are necessarily more stringent than those for models used purely for biological discovery.25

### 7.1 Validation Methods

A multi-faceted approach to validation is required, encompassing different stages and perspectives:

* **Internal Validation:** Assesses model performance using data withheld from the training process but drawn from the same source population. Common techniques include splitting data into training/testing sets or using k-fold cross-validation.51 This provides an initial check for overfitting but does not guarantee performance in new settings.
* **External Validation:** This is a crucial step for assessing generalizability. The model is tested on entirely separate datasets, ideally from different institutions, patient populations, geographical locations, or time periods.51 A lack of robust external validation is a frequently cited limitation in the ML-for-health literature, hindering confidence in model deployability.69 Consistent reporting of external validation efforts is needed.
* **Prospective Validation:** Evaluates model performance on new data as it becomes available, either in a silent/monitoring mode or within a simulated clinical workflow, before live deployment. This provides a more realistic assessment of performance in the target environment.
* **Comparison with Gold Standards:** Where feasible, model outputs can be compared against established benchmarks, such as expert clinician judgment, consensus guidelines, concurrently collected high-quality data (e.g., registry data, paper records in some contexts), or patient-reported information.64
* **Simulation-Based Validation:** For DTs involving simulation components, validation includes assessing the plausibility and accuracy of the simulation itself. For patient-specific DTs, this involves comparing the DT's predicted trajectory over time against the actual patient's observed clinical course, using new measurements to continually assess performance.26

### 7.2 Key Evaluation Metrics

Validation should encompass a range of metrics beyond simple accuracy:

* **Predictive Accuracy/Discrimination:** Measures the model's ability to correctly classify outcomes or predict values. Standard metrics include Area Under the Receiver Operating Characteristic curve (AUC or AUROC), Area Under the Precision-Recall curve (PR-AUC, often more informative for imbalanced datasets), sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, overall accuracy, F1-score, and Mean Absolute Error (MAE) for continuous predictions.21
* **Calibration:** Assesses how well the model's predicted probabilities align with the actual observed frequencies of the outcome. A well-calibrated model's prediction of "70% risk" should correspond to the event occurring in approximately 70% of patients assigned that risk. Poor calibration can render a model clinically misleading, even if it has high discrimination (AUC).62 Calibration is often underreported in validation studies, hindering reliability assessment.69 Dynamic calibration techniques may be needed for adaptive models.62
* **Robustness:** Evaluates the stability and consistency of model performance across different patient subgroups (e.g., defined by demographics, comorbidities), variations in input data quality, or slight changes in the clinical environment.41 The goal is to avoid "brittle" models that perform well in development but fail unexpectedly in real-world settings.41
* **Reproducibility:** Concerns whether the reported results can be independently replicated using the same data and methods. This requires transparent reporting of methodology, code availability, and data access protocols.9
* **Generalizability (Transportability):** Measures how well the model maintains its performance when applied to different patient populations, clinical settings, EHR systems, or geographical locations.14 External validation is the primary means of assessing generalizability. Limited validation across diverse international populations is a common gap.69
* **Uncertainty Quantification:** Explicitly measuring and communicating the model's uncertainty in its predictions is vital for safe clinical use.25 This includes distinguishing between *epistemic* uncertainty (due to limited data or model knowledge, potentially reducible) and *aleatoric* uncertainty (due to inherent randomness or noise in the system). Techniques like Bayesian inference, deep ensembles, Monte Carlo dropout, and entropy calculations can be used.62
* **Privacy Evaluation:** For models involving sensitive data or synthetic data generation, specific metrics assess the risk of re-identification or information leakage. Methods include calculating exact match rates, using shadow models, distance-based metrics, or assessing discriminator performance in GANs.66

Validation is clearly a multidimensional endeavor, requiring assessment across accuracy, calibration, robustness, generalizability, uncertainty, and potentially privacy. The relative importance of each dimension depends heavily on the DT's intended clinical application. For example, a DT used for risk stratification to guide treatment decisions requires excellent calibration, whereas a DT used for early alerting might prioritize sensitivity. The lack of standardized validation metrics and inconsistent reporting currently makes it difficult to compare models and reliably assess clinical readiness.69 Adopting standardized reporting guidelines (analogous to CONSORT for trials or TRIPOD for prediction models) and utilizing assessment frameworks like PROBAST 68 or QUADAS-2 68 could significantly improve the quality and comparability of validation studies.

### 7.3 Assessing Translational Readiness and Clinical Utility

Beyond technical performance metrics, assessing translational readiness requires evaluating the DT's potential for real-world clinical impact:

* **Actionability:** Does the DT provide information that is genuinely useful and actionable for clinicians or patients in making decisions? 24
* **Workflow Integration:** Can the DT be seamlessly integrated into existing clinical workflows without causing undue burden or disruption? 26 Presenting information at the right time, to the right person, in the right format is key (the "5 rights" of CDS).65
* **Clinical Utility:** Does using the DT actually lead to improved clinical decision-making, enhanced operational efficiency, better patient outcomes, or reduced costs? 41 As noted previously, demonstrating clinical utility through prospective studies remains a significant gap for many proposed DT applications. Decision Curve Analysis (DCA) can offer a preliminary assessment of potential clinical value by weighing the benefits of true positives against the harms of false positives across different risk thresholds.51

A specific challenge arises when validating DTs designed to be dynamic and adaptive, continuously learning from new data.4 Traditional validation methods often rely on static datasets and assume a fixed model. Validating a system that is constantly evolving requires different approaches. It necessitates ongoing monitoring and evaluation *after* deployment, not just pre-market testing.65 Methodologies must assess the stability, reliability, and safety of the learning and adaptation process itself. Techniques such as continuous performance monitoring, detection of data or model drift 65, and longitudinal comparison of the DT's predictions against the patient's actual evolving trajectory 26 become essential components of the validation lifecycle for these advanced, adaptive twins.

## 8. Ethical, Legal, and Social Implications (ELSI)

The development and deployment of EHR-based digital twins, given their reliance on vast amounts of personal health data and their potential to influence high-stakes clinical decisions, inevitably raise profound ethical, legal, and social implications (ELSI).7 Proactive identification and management of these issues are crucial for responsible innovation and public trust. ELSI research, supported by programs like those at the National Human Genome Research Institute (NHGRI) 72, plays a vital role in navigating this complex landscape. Training researchers and practitioners on digital health ELSI is also increasingly recognized as essential.74

### 8.1 Key ELSI Domains

Several key areas of ELSI concern are particularly relevant to healthcare DTs:

* **Patient Privacy and Data Security:** The aggregation and analysis of extensive, sensitive health data (EHRs, genomics, wearables, etc.) inherent to DTs create significant privacy risks.7 Ensuring data confidentiality, preventing unauthorized access or breaches, and complying with regulations like HIPAA and GDPR are non-negotiable.7 Robust technical measures (e.g., encryption, secure storage, access controls) and strong governance policies are required.7 The difficulty of achieving true anonymity with rich datasets further complicates privacy protection.67
* **Algorithmic Bias and Fairness:** DT models trained on historical EHR data may inherit and potentially amplify existing societal biases reflected in healthcare delivery and documentation.9 This can lead to DTs performing differently or providing inequitable recommendations for various demographic groups (e.g., based on race, ethnicity, gender, socioeconomic status), exacerbating health disparities.43 Addressing this requires methods for bias detection, mitigation strategies (e.g., fairness-aware algorithms, representative data collection), and ongoing audits for fairness throughout the DT lifecycle.9 Ensuring equitable benefits across diverse populations is a core goal.26
* **Informed Consent:** Obtaining truly informed consent for the complex, dynamic, and potentially long-term use of diverse data in a DT framework presents unique challenges.25 Patients need to understand how their data will be used, the potential risks and benefits, limitations of the technology, and how predictions are generated. Standard consent forms may be inadequate. Issues arise with secondary data use, the integration of continuously streaming data (e.g., from wearables), and data that may pertain to multiple individuals (e.g., genetic data).67 Dynamic consent models, allowing individuals ongoing control over permissions, have been proposed as a potential solution.67 Clear communication, managing expectations, and assessing understanding are critical.67
* **Autonomy and Patient Participation:** While DTs can potentially empower patients with more information 15, there's a risk of undermining patient autonomy if recommendations are overly prescriptive or poorly understood.2 Balancing technological guidance with individual choice and values is essential. The role of patient-generated health data needs careful consideration regarding accuracy and integration.67 Efforts are needed to prevent patient disenfranchisement, unjustified anxiety, or over-reliance on technology, particularly in the context of varying levels of digital health literacy.67 Enhancing patient engagement and education regarding their DT is important.15
* **Accountability and Liability:** Determining who is responsible when a DT provides incorrect predictions or recommendations leading to patient harm is complex.65 Is it the clinician who acted on the information, the developers of the DT algorithm, the institution deploying the technology, or the data providers? Lack of transparency in complex models ('black box' problem) further complicates accountability.9 Clear lines of responsibility and mechanisms for redress are needed. The opacity of some advanced models, particularly those using deep learning or LLMs, significantly amplifies these ethical concerns. If the reasoning behind a DT's output cannot be adequately explained, it erodes trust, hinders the detection of bias, complicates the process of obtaining informed consent, and makes assigning accountability extremely difficult. This underscores the critical need for research and implementation of explainable AI (XAI) techniques suitable for complex DTs, as well as robust methods for communicating model uncertainty.62
* **Data Ownership and Access:** Questions surrounding who owns the digital twin itself and the vast amounts of integrated patient data remain largely unresolved.25 Who has the right to access, modify, or delete the DT or its data? These issues require clear policy frameworks.
* **Resource Allocation and Equity:** The development and implementation of sophisticated DT technology require significant resources (financial, computational, expertise).9 There is a risk that DTs could exacerbate existing health disparities if they are primarily accessible only to well-resourced institutions or patient populations, creating a "digital divide" in access to advanced personalized medicine.16 Ensuring equitable access and benefit is a major societal challenge.

### 8.2 Regulatory Considerations

The regulatory landscape for AI/ML-based medical technologies, including DTs, is still evolving.7 Regulatory bodies like the U.S. Food and Drug Administration (FDA) are developing frameworks for evaluating the safety and effectiveness of AI/ML software as a medical device (SaMD).65 Key considerations include the level of autonomy of the device, whether it informs or drives clinical management, and the risk associated with its use.65 Clear regulatory pathways specifically addressing the unique characteristics of DTs (e.g., continuous learning, multi-component nature) will be necessary to facilitate safe and effective clinical translation.27

Addressing these multifaceted ELSI challenges cannot be an afterthought; it demands integration throughout the entire DT lifecycle, from initial design and data collection through development, validation, deployment, and ongoing monitoring.65 This necessitates a shift towards ethical design principles, incorporating privacy-by-design, conducting fairness audits, ensuring transparency, and actively engaging diverse stakeholders, including ethicists, social scientists, clinicians, patients, and policymakers, alongside engineers and data scientists.72 Institutional and funding support for such interdisciplinary collaboration and integrated ELSI research is paramount for building trustworthy and beneficial healthcare DTs.

## 9. Current Limitations and Future Frontiers

The field of EHR-based digital twins incorporating mechanistic models holds immense transformative potential, yet it remains in its early stages 4, facing significant limitations that must be overcome to realize its promise. Concurrently, exciting frontiers for research and development are emerging.

### 9.1 Recap of Current Limitations

The preceding sections have highlighted numerous challenges, which collectively represent the current limitations of the field:

* **Data Ecosystem Deficiencies:** Persistent issues with EHR data quality (accuracy, completeness, consistency), lack of interoperability between systems, difficulties in integrating heterogeneous data types, inherent biases reflecting care disparities, and stringent privacy/security constraints remain major roadblocks.7 The sheer volume and multi-modal nature of data required for high-fidelity twins often exceed practical availability.28
* **Modeling Challenges:** Accurately modeling the complexity, non-linearity, emergence, and heterogeneity of human biological systems remains profoundly difficult.14 Balancing the desire for mechanistic interpretability with the capabilities of data-driven methods, bridging multiple biological scales, and managing computational costs are ongoing struggles.14 Gaps in fundamental biological knowledge limit purely mechanistic approaches for many conditions.25
* **Validation and Clinical Utility Gaps:** Robust validation, particularly external and prospective validation demonstrating real-world effectiveness and generalizability, is often lacking.14 Standardized validation metrics and methodologies are needed, especially for dynamic, adaptive twins. Issues with calibration and robustness assessment persist.62 Critically, the translation from predictive accuracy to demonstrated clinical utility (improved decisions and outcomes) is frequently unproven.
* **ELSI and Regulatory Hurdles:** Numerous ethical, legal, and social issues surrounding privacy, bias, fairness, consent, accountability, and equity remain unresolved.7 Clear governance structures and regulatory pathways tailored to DTs are still under development.
* **Implementation Barriers:** Practical challenges include the high cost of development and implementation, the need for significant computational infrastructure, difficulties integrating DTs into complex clinical workflows, and achieving clinician buy-in and trust.9 Ensuring scalability across diverse healthcare settings is also a major concern.16

### 9.2 Future Research Directions and Frontiers

Despite the limitations, the field is rapidly advancing, with several key frontiers shaping future research:

* **Seamless Multi-Modal Data Integration:** Developing sophisticated methods and platforms to effectively integrate the full spectrum of relevant data – EHRs (structured/unstructured), genomics/multi-omics, imaging, real-time wearable/sensor data, patient-reported outcomes, SDoH – into coherent, dynamic patient models is a primary goal.4
* **Advancements in Algorithms:** Continued innovation in AI/ML is expected, focusing on developing more robust, interpretable (explainable AI), fair, and causally informative algorithms tailored for complex health data.4 Further exploration of LLMs for DT generation and interaction, addressing their limitations, is a key area.22 The potential long-term impact of quantum computing on simulation and analysis is also noted.36
* **Truly Dynamic and Adaptive Twins:** The vision extends to creating DTs that are not static snapshots but continuously learning, self-adapting systems ("intelligent," "cognitive," or "self-adaptive" twins) that evolve in real-time based on new patient data and feedback from clinical outcomes.4 This enables lifelong health monitoring and dynamic adjustment of predictions and recommendations.41
* **Enhanced Mechanistic and Multi-Scale Modeling:** Progress requires both deeper fundamental understanding of human biology and improved computational techniques for building and linking mechanistic models across multiple biological scales.25 Integrating hypothesis-driven, generative models with data-driven approaches is seen as essential for clinical relevance.28
* **Robust Validation Frameworks:** Developing and standardizing rigorous methodologies for validating complex, dynamic, and adaptive DTs is critical. This includes establishing best practices for assessing accuracy, calibration, robustness, generalizability, uncertainty, fairness, privacy, and, crucially, clinical utility.16 Utilizing DTs for *in silico* clinical trials is another avenue requiring validation.27
* **Mature ELSI Frameworks and Governance:** Establishing clear ethical guidelines, legal precedents, regulatory pathways, and governance structures specifically for healthcare DTs is essential for responsible development and deployment.65 This requires ongoing dialogue and collaboration between technologists, clinicians, ethicists, policymakers, and the public.
* **Interoperability Standards and Infrastructure:** Widespread adoption necessitates the development and enforcement of robust interoperability standards (for data formats, APIs like FHIR, clinical decision support integration like CDS Hooks) and the creation of scalable, secure computational and data infrastructure capable of supporting DT operations.7

The most promising trajectory likely involves the development of sophisticated **hybrid digital twins**. These systems will move beyond relying solely on data-driven patterns or potentially incomplete mechanistic knowledge. Instead, they will intelligently integrate (1) rich, multi-modal patient data streams, (2) explicit biological and mechanistic knowledge captured in computational models, (3) advanced AI/ML algorithms for learning, prediction, and adaptation, and (4) interactive interfaces allowing for clinician oversight, input, and potentially patient engagement.2 The future lies not in choosing between data-driven and knowledge-driven approaches, but in skillfully weaving them together, using real-world data to ground and personalize mechanistic components, and using mechanistic knowledge to structure analysis, enhance interpretability, and enable robust prediction beyond historical patterns. This necessitates flexible, modular DT architectures capable of integrating diverse components, with a strong emphasis on explainability and human-in-the-loop interaction.

Ultimately, the vision for healthcare DTs extends beyond passive prediction towards enabling **proactive and participatory healthcare**. Impactful DTs will not merely forecast risk but will empower clinicians and patients to simulate the effects of preventative strategies, monitor progress towards health goals in real-time, and collaboratively manage an individual's health trajectory over their lifespan.2 Achieving this requires not only accurate predictive models but also effective human-computer interfaces, tools for simulating interventions, integration with behavioral support systems, and careful navigation of the ethical considerations surrounding patient agency and guidance.

## 10. Conclusion

The integration of mechanistic models with Electronic Health Record data to create predictive digital twins represents a compelling frontier in healthcare informatics and personalized medicine. Originating from industrial applications, the healthcare DT concept leverages the convergence of ubiquitous EHRs, advanced AI/ML, multi-modal data streams, and increasing computational power to build dynamic, individualized virtual representations of patients or health systems. These DTs, distinguished from traditional simulations and predictive models by their persistent nature and continuous, bidirectional data connection to their real-world counterparts, aim to forecast disease onset, predict progression, optimize treatment selection, and identify risks with unprecedented granularity.

Incorporating mechanistic understanding – knowledge of underlying biological pathways and physiological processes – is crucial for moving beyond correlation to causation, enhancing model interpretability, and enabling the simulation of novel interventions. Hybrid approaches combining the pattern-recognition strengths of ML/AI with the explanatory power of mechanistic models appear most promising, though significant challenges remain in bridging the gap between complex biological reality and the often limited, noisy data available in EHRs. Case studies, particularly in fields like psychiatry, oncology, and cardiology, demonstrate the potential of these approaches, while also highlighting the critical role of NLP in extracting vital information from unstructured clinical notes and the persistent gap between achieving predictive accuracy and demonstrating tangible improvements in clinical outcomes.

Realizing the full potential of EHR-based mechanistic DTs requires addressing substantial hurdles. These include systemic data challenges related to quality, interoperability, bias, and privacy; the inherent difficulties in modeling complex biological systems; the need for rigorous, multi-dimensional validation frameworks suitable for dynamic, adaptive systems; and the critical imperative to navigate the complex ethical, legal, and social implications surrounding data ownership, fairness, consent, and accountability.

Future progress hinges on continued innovation across multiple fronts: developing advanced algorithms (including explainable AI and causal inference methods), seamlessly integrating diverse multi-modal data, creating truly dynamic and adaptive learning twins, enhancing mechanistic modeling capabilities, establishing robust validation standards and ELSI governance, and building the necessary interoperable infrastructure. The path forward necessitates strong interdisciplinary collaboration between data scientists, computational modelers, clinicians, biologists, ethicists, social scientists, and patients.

While significant research, development, and societal deliberation are still required, the vision of using EHR-based, mechanistically informed digital twins to enable proactive, personalized, and participatory healthcare offers a powerful motivation for continued effort in this transformative field.

#### Works cited

1. Digital twins for chronic lung diseases - PMC - PubMed Central, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11653195/>
2. The Digital Twin in Medicine: A Key to the Future of Healthcare? - PMC - PubMed Central, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9330225/>
3. Digital Twins for Clinical and Operational Decision-Making: Scoping Review - PMC, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11754991/>
4. Medical Digital Twin: A Review on Technical Principles and Clinical Applications - MDPI, accessed April 29, 2025, <https://www.mdpi.com/2077-0383/14/2/324>
5. Digital twin for healthcare systems - PMC, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10513171/>
6. Digital Twins' Advancements and Applications in Healthcare, Towards Precision Medicine - PMC - PubMed Central, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11595921/>
7. Digital twins for health: a scoping review - PMC, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10960047/>
8. (PDF) Digital Twin Technology in Healthcare: Applications, Challenges, and Future Insights, accessed April 29, 2025, <https://www.researchgate.net/publication/387504294_Digital_Twin_Technology_in_Healthcare_Applications_Challenges_and_Future_Insights>
9. Advancing Healthcare Through the Integration of Digital Twins Technology: Personalized Medicine's Next Frontier - MDPI, accessed April 29, 2025, <https://www.mdpi.com/1999-5903/16/12/477>
10. A Conceptual Framework for Digital Twin in Healthcare: Evidence from a Systematic Meta-Review - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/383978368_A_Conceptual_Framework_for_Digital_Twin_in_Healthcare_Evidence_from_a_Systematic_Meta-Review>
11. Envisioning the Future of Personalized Medicine: Role and Realities of Digital Twins - PMC, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11130780/>
12. Digital Twins in Healthcare: A Survey of Current Methods - Fortune Journals, accessed April 29, 2025, <https://fortuneonline.org/articles/digital-twins-in-healthcare-a-survey-of-current-methods.html>
13. Digital twins for health: a scoping review - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/379208495_Digital_twins_for_health_a_scoping_review>
14. Digital twin in healthcare: Recent updates and challenges - PMC, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9830576/>
15. Digital Twins in Healthcare: Simulating Patient Conditions for Better Outcomes, accessed April 29, 2025, <https://www.holonsolutions.com/digital-twins-in-healthcare-simulating-patient-conditions-for-better-outcomes/>
16. Advancing Health Care With Digital Twins: Meta-Review of ..., accessed April 29, 2025, <https://www.jmir.org/2025/1/e69544>
17. Advancing Health Care With Digital Twins: Meta-Review of Applications and Implementation Challenges, accessed April 29, 2025, <https://www.jmir.org/2025/1/e69544/PDF>
18. (PDF) Revolutionizing Healthcare: A Review Unveiling the Transformative Power of Digital Twins - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/380550890_Revolutionizing_Healthcare_A_Review_Unveiling_the_Transformative_Power_of_Digital_Twins>
19. Digital Twins Generated by Artificial Intelligence in Personalized Healthcare - MDPI, accessed April 29, 2025, <https://www.mdpi.com/2076-3417/14/20/9404>
20. Envisioning the Future of Personalized Medicine: Role and Realities of Digital Twins, accessed April 29, 2025, <https://www.jmir.org/2024/1/e50204/>
21. The Role of Digital Twins in Optimizing Healthcare Coverage Strategies - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/388836567_The_Role_of_Digital_Twins_in_Optimizing_Healthcare_Coverage_Strategies>
22. Large Language Models forecast Patient Health Trajectories enabling Digital Twins, accessed April 29, 2025, <https://www.medrxiv.org/content/10.1101/2024.07.05.24309957v1.full-text>
23. The Emerging Landscape of Health Research Based on Biobanks Linked to Electronic Health Records - Deep Blue Repositories, accessed April 29, 2025, <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/154448/sim8445.pdf?sequence=2>
24. Toward mechanistic medical digital twins: some use cases in immunology - PubMed, accessed April 29, 2025, <https://pubmed.ncbi.nlm.nih.gov/38515550/>
25. Toward mechanistic medical digital twins: some use cases in ..., accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10955144/>
26. Digital Twins for Predictive Oncology will be a Paradigm Shift for ..., accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9097784/>
27. From bits to bedside: entering the age of digital twins in cardiac ..., accessed April 29, 2025, <https://academic.oup.com/europace/article/26/12/euae295/7926243>
28. Challenges and opportunities for digital twins in precision medicine: a complex systems perspective - arXiv, accessed April 29, 2025, <https://arxiv.org/html/2405.09649v1>
29. Challenges and opportunities for digital twins in precision medicine from a complex systems perspective - PubMed, accessed April 29, 2025, <https://pubmed.ncbi.nlm.nih.gov/39825012/>
30. Challenges and opportunities for digital twins in precision medicine from a complex systems perspective - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/388114616_Challenges_and_opportunities_for_digital_twins_in_precision_medicine_from_a_complex_systems_perspective>
31. [2405.09649] Challenges and opportunities for digital twins in precision medicine: a complex systems perspective - arXiv, accessed April 29, 2025, <https://arxiv.org/abs/2405.09649>
32. Challenges and opportunities for digital twins in precision medicine: a complex systems perspective - arXiv, accessed April 29, 2025, <https://arxiv.org/pdf/2405.09649>
33. Toward mechanistic medical digital twins: some use cases in immunology - Frontiers, accessed April 29, 2025, <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1349595/full>
34. Advancing Clinical Information Systems: Harnessing Telemedicine, Data Science, and AI for Enhanced and More Precise Healthcare Delivery - PMC - PubMed Central, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12020641/>
35. Advancing Precision Oncology with Digital and Virtual Twins: A Scoping Review - MDPI, accessed April 29, 2025, <https://www.mdpi.com/2072-6694/16/22/3817>
36. Concepts and applications of digital twins in healthcare and medicine - Sheng Xu, accessed April 29, 2025, <https://xugroup.eng.ucsd.edu/wp-content/uploads/2024/11/94.pdf>
37. Transforming Precision Medicine - MDPI, accessed April 29, 2025, <https://mdpi-res.com/bookfiles/book/9150/Transforming_Precision_Medicine_The_Intersection_of_Digital_Health_and_AI.pdf?v=1744765587>
38. Building Digital Twins for Cardiovascular Health: From Principles to Clinical Impact | Journal of the American Heart Association, accessed April 29, 2025, <https://www.ahajournals.org/doi/10.1161/JAHA.123.031981>
39. Digital Twins' Advancements and Applications in Healthcare, Towards Precision Medicine, accessed April 29, 2025, <https://www.mdpi.com/2075-4426/14/11/1101>
40. Digital Twins for Health: A Scoping Review - DigitalCommons@TMC, accessed April 29, 2025, <https://digitalcommons.library.tmc.edu/cgi/viewcontent.cgi?article=1514&context=uthshis_docs>
41. How Digital Twin Technology Can Revolutionize Healthcare ..., accessed April 29, 2025, <https://siim.org/artificial-intelligence/how-digital-twin-technology-can-revolutionize-healthcare/>
42. Digital twins in healthcare: practical lessons from a Canadian hospital - AnyLogic, accessed April 29, 2025, <https://www.anylogic.com/blog/digital-twins-in-healthcare-practical-lessons-from-a-canadian-hospital/>
43. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm, accessed April 29, 2025, <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2018.00031/full>
44. Simulation vs Digital Twin (What is the Difference Between Them ..., accessed April 29, 2025, <https://www.twi-global.com/technical-knowledge/faqs/simulation-vs-digital-twin>
45. Foundational Research Gaps and Future Directions for Digital Twins - PubMed, accessed April 29, 2025, <https://pubmed.ncbi.nlm.nih.gov/39088664/>
46. TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model - arXiv, accessed April 29, 2025, <https://arxiv.org/html/2404.01273v2>
47. TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model - arXiv, accessed April 29, 2025, <https://arxiv.org/pdf/2404.01273>
48. Digital Twin vs. Simulation: Key Differences - Program-Ace, accessed April 29, 2025, <https://program-ace.com/blog/digital-twin-vs-simulation/>
49. Patient twin: Why Sophia is no longer afraid of cancer - Siemens Healthineers, accessed April 29, 2025, <https://www.siemens-healthineers.com/perspectives/digital-patient-twin-cancer-care>
50. Causal machine learning for healthcare and precision medicine - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/362454203_Causal_machine_learning_for_healthcare_and_precision_medicine>
51. Applying contrastive pre-training for depression and anxiety risk prediction in type 2 diabetes patients based on heterogeneous electronic health records: a primary healthcare case study - Oxford Academic, accessed April 29, 2025, <https://academic.oup.com/jamia/article/31/2/445/7462201>
52. Toward personalizing treatment for depression: predicting diagnosis and severity | Journal of the American Medical Informatics Association | Oxford Academic, accessed April 29, 2025, <https://academic.oup.com/jamia/article/21/6/1069/2909304>
53. (PDF) Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/350130469_Natural_language_processing_and_machine_learning_of_electronic_health_records_for_prediction_of_first-time_suicide_attempts>
54. Scientists Develop a 'Digital Twin' Model to Predict Cancer Treatment Responses, accessed April 29, 2025, <https://ascopost.com/news/october-2024/scientists-develop-a-digital-twin-model-to-predict-cancer-treatment-responses/>
55. High throughput phenotyping for dimensional psychopathology in electronic health records, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC5972065/>
56. Search AMIA.org, accessed April 29, 2025, [https://amia.org/search?f%5B0%5D=type%3AGroups%20%20Communities&f%5B1%5D=type%3AMeetings&f%5B2%5D=type%3AOther&f%5B3%5D=ujournal%3AJAMIA&f%5B4%5D=ujournal%3AJAMIA%20Open&page=13](https://amia.org/search?f%5B0%5D=type:Groups++Communities&f%5B1%5D=type:Meetings&f%5B2%5D=type:Other&f%5B3%5D=ujournal:JAMIA&f%5B4%5D=ujournal:JAMIA+Open&page=13)
57. Automated detection of substance use information from electronic health records for a pediatric population, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8449626/>
58. Applications of Natural Language Processing in the Domain of Mental Health, accessed April 29, 2025, <https://www.researchgate.net/publication/385313771_Applications_of_Natural_Language_Processing_in_the_Domain_of_Mental_Health>
59. Implementation of Custom Prediction Models in the EHR - A Case ..., accessed April 29, 2025, <https://amia.org/webinar-library/implementation-custom-prediction-models-ehr-case-study-postpartum-depression>
60. Digital Twin Ecosystem for Oncology Clinical Operations - arXiv, accessed April 29, 2025, <https://arxiv.org/html/2409.17650v1>
61. Cardiac Healthcare Digital Twins Supported by Artificial Intelligence-Based Algorithms and Extended Reality—A Systematic Review - MDPI, accessed April 29, 2025, <https://www.mdpi.com/2079-9292/13/5/866>
62. The challenge of uncertainty quantification of large language models in medicine - arXiv, accessed April 29, 2025, <https://arxiv.org/html/2504.05278v1>
63. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review - PMC - PubMed Central, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7671639/>
64. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3555312/>
65. Regulatory Aspects and Ethical Legal Societal Implications (ELSI) - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/378732057_Regulatory_Aspects_and_Ethical_Legal_Societal_Implications_ELSI>
66. Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics, accessed April 29, 2025, <https://www.medrxiv.org/content/10.1101/2023.11.28.23299124v1.full-text>
67. Digital Technologies and Data Science as Health Enablers: An Outline of Appealing Promises and Compelling Ethical, Legal, and Social Challenges - Frontiers, accessed April 29, 2025, <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.647897/epub>
68. Use of Machine Learning and Artificial Intelligence Methods in Geriatric Mental Health Research Involving Electronic Health Record or Administrative Claims Data: A Systematic Review - Frontiers, accessed April 29, 2025, <https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2021.738466/full>
69. (PDF) Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: a scoping review - ResearchGate, accessed April 29, 2025, <https://www.researchgate.net/publication/389215596_Externally_validated_and_clinically_useful_machine_learning_algorithms_to_support_patient-related_decision-making_in_oncology_a_scoping_review>
70. Reinforcement Learning in Personalized Medicine: A Comprehensive Review of Treatment Optimization Strategies | Cureus, accessed April 29, 2025, [https://www.cureus.com/articles/359834-reinforcement-learning-in-personalized-medicine-a-comprehensive-review-of-treatment-optimization-strategies.pdf?email=](https://www.cureus.com/articles/359834-reinforcement-learning-in-personalized-medicine-a-comprehensive-review-of-treatment-optimization-strategies.pdf?email)
71. AMIA 2024 Annual Symposium On Demand, accessed April 29, 2025, <https://amia.org/education-events/amia-2024-annual-symposium-demand>
72. Ethical, Legal and Social Implications Research Program, accessed April 29, 2025, <https://www.genome.gov/Funded-Programs-Projects/ELSI-Research-Program-ethical-legal-social-implications>
73. Ethical, Legal and Social Implications (ELSI) Research Advisors Report to the National Advisory Council for Human Genome Resear - ELSIhub, accessed April 29, 2025, <https://elsihub.org/sites/default/files/content-files/ERAReportdraft5.pdf>
74. Ethical, legal, and social implications of digital health: A needs assessment from the Society of Behavioral Medicine to inform capacity building for behavioral scientists - PubMed Central, accessed April 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10890818/>
75. Ethical, legal, and social implications of digital health: A needs assessment from the Society of Behavioral Medicine to inform capacity building for behavioral scientists - PubMed, accessed April 29, 2025, <https://pubmed.ncbi.nlm.nih.gov/38011809/>