Therapeutic Innovation & Regulatory Science
https://doi.org/10.1007/s43441-026-01067-y

ANALYSIS

The Role of Digital Twins in Clinical Trials: A New Perspective on Methodological and Statistical Challenges

Fatma Nur Cicin[1]

Received: 17 March 2026 / Accepted: 16 October 2026
© The Author(s), under exclusive licence to The Drug Information Association, Inc 2026

Abstract

Background Randomized controlled trials (RCTs) remain the gold standard for evaluating medical interventions, yet they face persistent methodological challenges: patient dropout, inadequate sample sizes, premature termination, missing data, and the limitations of indirect treatment comparisons. The emergence of digital twin technology—computational patient representations that generate simulated clinical trajectories statistically indistinguishable from real-world data—offers a promising methodological complement to address these limitations.

Approach We synthesize evidence from key methodological literature, regulatory guidance (ICH E9(R1), FDA, EMA), and proof-of-concept studies across oncology, neurology, and Alzheimer's disease. A taxonomy of in silico clinical trial approaches provides the organizational framework. This article is a narrative perspective rather than a systematic review; no predefined search strategy, database range, or inclusion/exclusion criteria were applied.

Results Digital twins may provide complementary methodological approaches across nine trial challenges: model-based sensitivity analyses for missing (MNAR) outcomes; Bayesian and posterior predictive modeling of partially observed longitudinal outcomes; prognostic covariate adjustment and virtual augmentation to improve efficiency; prospective simulation of futility stopping rules; prospective calibration of efficacy stopping boundaries; individualized control-arm (ICA) construction; hybrid augmented and synthetic control strategies; pre-trial protocol optimization through in silico simulation; and population-adjusted indirect treatment comparison.

Conclusions Digital twins represent an emerging methodological ecosystem rather than isolated tools. Their methodological potential depends on consistent validation and explainability standards; frameworks for detecting algorithmic bias; clarification of regulatory acceptance criteria; and comprehensive data infrastructure addressing representativeness gaps. Digital twin approaches should be viewed as complementary to—not replacements for—established randomized trial methodology.

Keywords Digital twin · Clinical trial methodology · Missing data · Statistical power · External control arm · Population-adjusted indirect comparison · Artificial intelligence · In silico trial

Introduction

Randomized controlled trials (RCTs) remain the methodological gold standard for evaluating the efficacy and safety of medical interventions. Nevertheless, the challenges encountered in the design and conduct of clinical trials — patient dropout, slow enrollment, premature termination, missing data, and the limitations of indirect comparisons— pose serious threats to the statistical validity and clinical generalizability of their findings. This process has become increasingly time-consuming and costly: the median cost of a Phase I trial is $3.4 million, while a Phase III trial exceeds $21.4 million. Eighty percent of clinical trials experience enrollment delays, 20% fail to meet their target sample size, and only 10–13% of drugs progress from Phase I to regulatory approval [1, 2].

In parallel with these challenges, advances in artificial intelligence (AI), machine learning, and large-scale biomedical data integration have introduced the concept of the digital twin into biomedical research. The concept

✉ Fatma Nur Cicin
fatma.cicin@istinye.edu.tr

[1] Department of Management Information Systems, Faculty of Economics, Administrative and Social Sciences, Istinye University, Ayazağa Mah, Azerbaycan Cad, (Vadistanbul 4A Blok), Sarıyer, 34396 Istanbul, Türkiye

Published online: 23 October 2026

was originally proposed by Grieves in the context of product lifecycle management and later adopted in aerospace engineering by NASA as a computational framework for simulating complex physical systems [3]. Recent scoping reviews confirm a rapid surge in healthcare digital twin research across disease management, clinical decision support, and trial design [4, 5]. Over the past decade, the digital twin paradigm has increasingly been explored in biomedical research as a means of modeling complex biological systems and predicting disease trajectories using multimodal patient data [2].

In healthcare, a digital twin can be broadly defined as a computational representation of an individual patient capable of generating simulated clinical trajectories that are statistically indistinguishable from real-world outcomes [1]. These models typically integrate heterogeneous data sources such as genomic information, imaging data, electronic health records, and patient-reported outcomes to simulate disease progression and treatment response. Advances in generative modeling approaches — including variational autoencoders, generative adversarial networks, and probabilistic neural networks — have further expanded the potential of digital twins to simulate individual-level counterfactual outcomes [2].

Recent research has suggested that digital twins may have important implications for clinical trial methodology. Digital twin models have been proposed as tools for modeling missing data trajectories, improving statistical efficiency through prognostic covariate adjustment, and constructing individualized external control arms when randomization is difficult or ethically problematic [1, 6]. Simulation-based approaches may further support clinical trial design optimization through in silico experiments evaluating recruitment scenarios, endpoint definitions, and adaptive design strategies [2].

In this perspective article, we examine nine major methodological challenges in contemporary clinical trials and explore how digital twin methodologies may contribute to addressing these issues. Particular attention is given to statistical challenges such as missing data mechanisms, insufficient sample sizes, early stopping decisions, hybrid control arm designs, and indirect treatment comparisons. Consistent with this aim, the present work is intended as a perspective / commentary and not as a systematic review; it does not apply a formal search protocol. Its principal original contribution is the explicit linkage, developed in "Indirect Treatment Comparisons and Digital Twin–Based Simulation" section, between individualized digital-twin simulation and population-adjusted indirect treatment comparison methods (MAIC, STC, and IPD-NMA); "Patient Dropout: Missing Data Mechanisms and Digital Twin Simulation"–"Hybrid Designs: Augmented and Synthetic Control" section are intended to consolidate, rather than to extend, concepts already described in recent reviews [1, 2, 4]. Throughout, we emphasize a distinction that is central to the correct use of these methods: a digital twin produces a prediction or simulation, and the prediction of an unobserved outcome is not equivalent to identification of that outcome, to reconstruction of a randomized comparator, or to estimation of an individual causal effect. Mechanistic simulation, prognostic prediction, generative and synthetic-data generation, prognostic covariate adjustment in randomized trials, external or synthetic control-arm construction, Bayesian borrowing from historical information, and patient-level counterfactual estimation rest on different assumptions, validation requirements, and regulatory implications, and should not be treated as interchangeable forms of evidence.

Digital Twins: Conceptual Framework, Historical Development, and Construction

Origins and Historical Background

The term ‘digital twin’ was first coined in 2002 by Michael Grieves of the University of Michigan in the context of product lifecycle management [3]. The concept was operationalized during NASA’s Apollo 13 mission, where a simulation of the spacecraft supported critical navigation decisions that enabled the safe return of the astronauts. In 2010, NASA formally defined the digital twin as ‘an integrated, multi-physics, multi-scale, probabilistic simulation of a vehicle or system that uses the best available physical models, sensor updates, and fleet history to mirror the life of its flying twin’ [1].

Following widespread adoption in aerospace and manufacturing under the Industry 4.0 paradigm, the technology entered the biomedical research agenda from approximately 2015 onward. AI applications have evolved from logic-based symbolic reasoning and linear regression models through deep learning-based neural networks to large general-purpose models built on transformer architectures [2].

Definition of AI-Generated Digital Twins

Mechanistic (physics-based) models have proven successful in constructing digital twins of individual organ systems: in cardiology, Dassault’s Living Heart project examines organ–drug interactions in silico, while FEops HEARTguide is used for planning cardiac procedures [2]. Vidovszky et al. define the AI-generated digital twin as ‘a virtual patient representation, constructed using modern AI methods applied to large clinical datasets, from which predicted trajectories statistically indistinguishable from the patient’s real data

can be generated.' Compared with mechanistic models that impose explicit structural assumptions, generative AI models may require fewer prespecified mechanistic constraints; however, their predictive performance remains dependent on training-data quality, calibration, and transportability [1]. A 2025 Lancet Digital Health framework further identifies five core architectural components of the medical digital twin: the patient, data connection, patient-in-silico, interface, and twin synchronization—providing a conceptual map for regulatory evaluation and clinical deployment [5].

A Taxonomy of In Silico Clinical Trials

A widely cited taxonomy organizes AI-based in silico clinical trials under three main headings [2, 4]:

Clinical simulation: Investigates the efficacy and safety of treatments by generating virtual cohorts or organisms, encompassing (a) digital twins—generative AI models (CRBM-, GAN-, VAE-based) producing control trajectories statistically indistinguishable from real data, and [b) biosimulations—mathematical modeling of physiology using differential equations and physics-based models. The FDA-approved virtual glucose control simulator and Dassault's Living Heart Project are widely cited examples in diabetes and cardiology [2, 7].

Individualized predictive modeling: Encompasses disease onset, mortality, and readmission prediction using gradient boosting trees, LSTM, and transformer models; and multi-label clinical event prediction from longitudinal EHR data [2].

Computer-aided trial design: Includes trial semantic parsing (extraction of PICO elements from free-text protocols via NLP), outcome prediction (trial success/failure forecasting using multi-modal real-world data), and protocol evaluation (computing the marginal contribution of eligibility criteria to cohort size and hazard ratios via Shapley scores) [2].

The overall architecture of a digital twin framework in clinical research is illustrated in Fig. 1.

Multimodal patient data—including multi-omics, medical imaging, electronic health records (EHR), and patient-reported outcomes (PROs)—are integrated and harmonized to train artificial intelligence and machine learning models. These models generate a digital twin that simulates disease trajectories and treatment responses, enabling outcome simulation and supporting clinical trial decision-making.

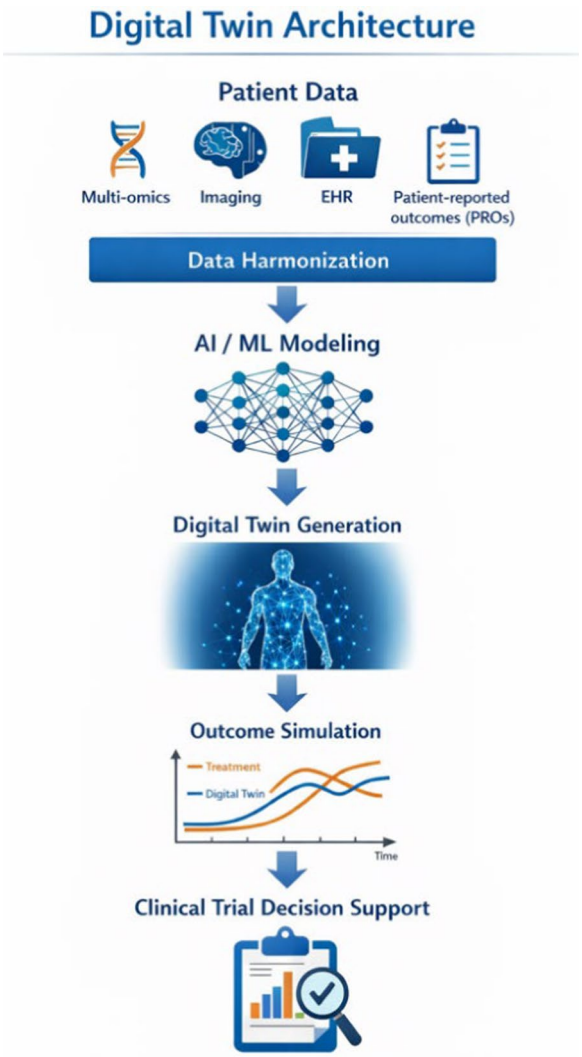


Fig. 1 Architecture of a digital twin framework in clinical research

Structural Components: Three Core Blocks

Mollica et al. describe the digital twin architecture through three interconnected blocks [8]:

Data collection block: Encompasses the harmonization and standardization of multi-modal patient data. FAIR principles, OMOP CDM, and FHIR provide the foundational framework [9]. EHRs offer rich data; however, format inconsistencies and poorly labeled disease definitions constrain AI models [1].

Modeling block: A disease-specific AI model trained on historical control data generates probabilistic control outcome predictions from each participant's pre-randomization baseline data. Data drift and changes in standard of care necessitate robust validation and update procedures [1, 8].

Connectivity block: Manages the transfer of heterogeneous input data to the digital twin and the delivery of outputs to the real world. IoT integration, real-time synchronization, and cybersecurity are critical dimensions. Transfer learning enables feature transfer from general-domain pretrained models to healthcare applications [8].

Current Clinical Applications

In oncology, Tardini et al. applied a deep Q-learning-based dual digital twin system to a dataset of 536 patients with oropharyngeal squamous cell carcinoma, achieving high accuracy in long-term outcome and toxicity prediction [10]. Chaudhuri et al. performed patient-specific radiotherapy dose optimization for high-grade glioma using digital twins with Bayesian model calibration [11]. Wu et al. predicted individual treatment responses in breast tumors using imaging-guided digital twins integrating multiple imaging modalities [12].

In neurology, research has demonstrated that probabilistic neural network-based digital twins can generate virtual patient trajectories statistically indistinguishable from real patients enrolled in the placebo arms of Multiple Sclerosis clinical trials — a finding with direct implications for virtual control arm design [13]. Computational modeling approaches have further shown that integrating digital twins into neurological trial design can substantially reduce required sample sizes. In Alzheimer's disease, unsupervised machine learning models have been developed that simulate comprehensive disease trajectories indistinguishable from real patient data, providing a foundation for in silico control arms [14, 15].

The Digital Twin Construction Process

The construction of a clinical research digital twin proceeds through four fundamental stages (Fig. 2):

Stage 1—*Data collection and harmonization:* EHR, omics, imaging, and PRO data are transferred to a standard data model. Open data sources such as MIMIC-III/IV and Project Data Sphere, and ClinicalTrials.gov with access to over 400,000 trial records, provide important training infrastructure [2].

Stage 2—*Model training:* A disease-specific AI model trained on historical control data generates probabilistic control outcome predictions from baseline data collected prior to randomization [1].

Stage 3—*Validation and explainability:* Discrimination, calibration, and clinical utility metrics are assessed in independent cohorts. SHAP quantifies each feature's contribution to individual predictions; LIME approximates black-box models with locally interpretable ones. Robustness to missing input data is documented via input sensitivity analysis [1].

Stage 4—*Simulation and updating:* The validated model is integrated into the simulation engine for study-specific scenarios. Through a Bayesian update mechanism, the model is revised as new patient data arrive. Data drift and the introduction of new biomarkers require continuous monitoring throughout the model's lifecycle [8].

Digital twin development begins with multimodal patient data collection and harmonization. Machine learning models are trained on these data to generate a virtual patient representation. Clinical outcomes are then used to iteratively update and refine the digital twin, improving predictive performance over time.

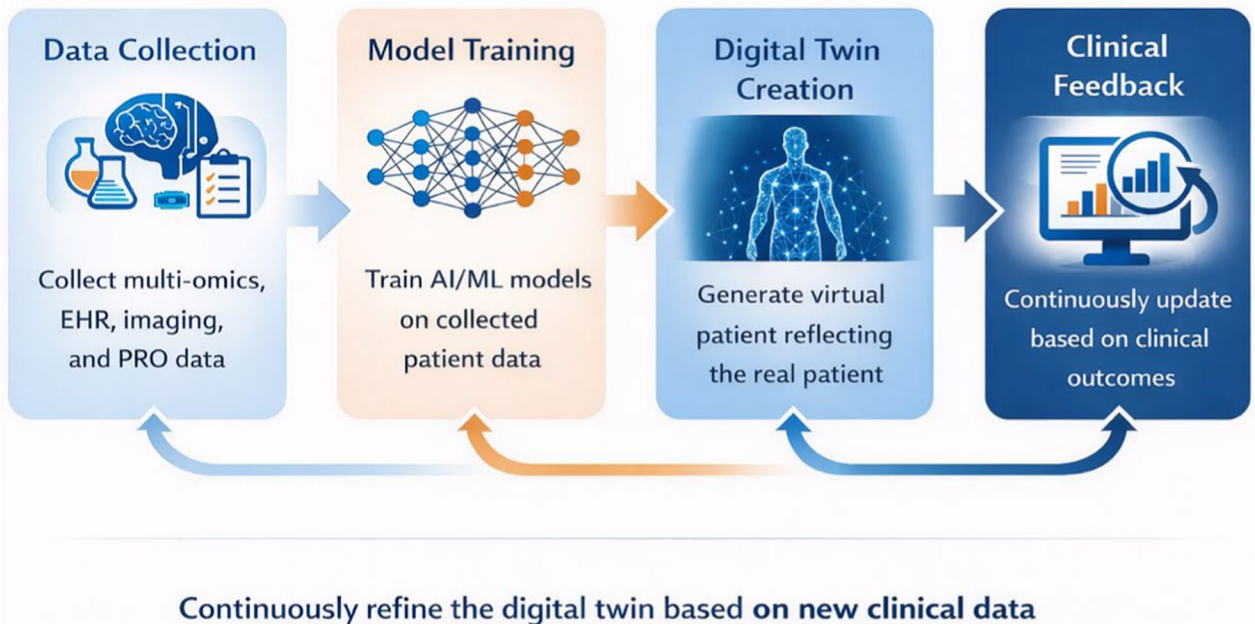


Fig. 2 The digital twin construction process

Core Challenges in Clinical Trials and Digital Twin Approaches

Patient Dropout: Missing Data Mechanisms and Digital Twin Simulation

Patient dropout is not merely a mechanical problem reducing sample size; it is a critical methodological variable that determines the direction of bias in results. The classical taxonomy of missing data mechanisms—formalized by Rubin and widely adopted in regulatory guidance—identifies three categories: under MCAR, the probability of missingness is unrelated to any observed or unobserved values; under MAR, missingness depends only on observed variables; under MNAR, the probability of missingness is related to the unobserved values themselves (e.g., patients who withdraw due to severe toxicity) [16, 17]. Under MAR, multiple imputation (MI; $m \geq 5$ imputation sets, combined via Rubin's rules) and EM algorithm-based maximum likelihood produce consistent and efficient estimates; under MNAR, pattern mixture models and selection models provide parametric sensitivity analyses, though their assumptions cannot be verified from the data [17].

ICH E9[R1] recommends structured sensitivity analyses (control-based imputation, reference-based imputation) in the MNAR context while excluding LOCF as a primary analysis method, owing to its unrealistic assumption that disease progression halts upon treatment discontinuation [18]. Digital twins offer a model-based alternative: for each patient who withdraws, an individualized computational model is constructed from their biomarker trajectory and clinical profile up to the point of withdrawal, replacing the external parametric assumptions of pattern-mixture or selection models with a different, and equally unverifiable, set of model-based assumptions (model specification, training-data representativeness, temporal transportability, measurement error, extrapolation beyond the observed trajectory, and the relationship between the modelled variables and the reasons for dropout). Robustness to missing input data must be documented via input sensitivity analysis [1]. Under an MNAR mechanism the distribution of the unobserved outcomes is generally not identifiable from the observed data without unverifiable assumptions, and a digital-twin model does not remove this nonidentifiability; it substitutes assumptions about model form and training data for the parametric assumptions of conventional sensitivity analyses. Digital-twin trajectories may generate clinically plausible post-dropout values, but these remain model-based predictions and should not be interpreted as revealing the patient's unobserved biological process. Consistent with ICH E9(R1), which emphasizes clear estimand definitions and sensitivity analyses rather than a predictive resolution of missingness, such models are best positioned within prespecified estimands and sensitivity analyses rather than as a solution to MNAR nonidentifiability.

Partial Participation and Bayesian Updating

Patients with partial participation—those who attend multiple assessment visits but withdraw before the final primary outcome assessment—are included in the ITT population yet contribute limited analytical power due to missing endpoint data; this problem is particularly pronounced in trials with long follow-up requirements, such as overall survival and progression-free survival studies. The Bayesian framework allows for the probabilistic updating of available evidence: observed patient data, including interim longitudinal measurements, contribute to the likelihood, while the prior represents information specified before observing the current data (for example historical studies or a separately trained prognostic model). A model-generated prediction of an unobserved endpoint does not by itself constitute a valid likelihood contribution, and using the same interim data both to generate a prediction and to treat that prediction as additional observed evidence risks the double use of information. Any such analysis must therefore propagate the uncertainty arising from model training, model selection, transportability, and imputation into the final inference, and must distinguish (a) Bayesian updating of model parameters from observed longitudinal data, (b) Bayesian borrowing from historical controls, (c) posterior predictive distributions for missing outcomes, and (d) prognostic covariate adjustment using an externally trained prediction model, which are related but fundamentally different procedures [13, 19]. A related but distinct application is the PROCOVA+™ framework. In rare diseases and pediatric populations, historical data can be converted into informative prior distributions to improve the accuracy and precision of treatment effect estimates in small-sample trials; sensitivity analyses regarding the degree of informativeness of the prior must be reported [6].

Inadequate Sample Size: Statistical Power and Digital Twin Strategies

Statistical power $(1-\beta)$ is the probability of rejecting the null hypothesis when a true effect exists, and constitutes one of the most critical parameters in clinical trial design. For a simplified two-group comparison of a continuous outcome under equal allocation, the required sample size per group is $n = 2\sigma 2(z_\alpha/2 + z_\beta)2 / \delta 2$. Three fundamental sources of uncertainty dominate this formula: [1] incorrect estimation of the effect size (δ) from prior studies; (2) insufficient knowledge of the outcome variable variance ($\sigma 2$); and (3)

actual dropout rates exceeding projections. Approximately 40% of oncology clinical trials are known to fail to achieve their target sample size (8).

Digital twin technology can contribute to this problem through two distinct mechanisms. The first is virtual cohort augmentation: digital patients generated from historical trial databases and real-world data — faithfully reflecting the statistical distributional properties of the real cohort — are added to increase effective sample size. Generating additional records does not, however, generate additional independent information: when synthetic patients are simulated from the same historical data and fitted model, their information content is bounded by that data, and treating them as independent observations can produce artificially narrow confidence or credible intervals and inflate the type I error rate. Virtual augmentation should therefore be accompanied by appropriate discounting (for example power-prior down-weighting) and by evaluation of operating characteristics. This approach provides a critical methodological alternative particularly in rare diseases and pediatric populations where conventional enrollment is not feasible [1, 8]. The second mechanism is *prognostic covariate control:* the control scenario predicted by each patient's digital twin is used as a covariate in the analysis, potentially reducing the residual variance of the treatment effect estimate when the prognostic model is well calibrated. This approach offers a means of increasing statistical power from existing data without augmenting sample size. Both mechanisms are particularly relevant in pediatric and rare disease trials, where ethical and practical constraints make conventional enrollment infeasible [20].

A concrete application operationalizing the power contribution of digital twins is the PROCOVA™ (Prognostic Covariate Adjustment) framework applied by Unlearn.ai to Alzheimer's disease clinical trials. Rather than using simple baseline covariates as in ANCOVA, this method uses the predicted outcome calculated from each patient's digital twin as the covariate. Schuler et al. [2022] demonstrated in a peer-reviewed publication that this procedure yields minimum variance unbiased estimators, and it has been recognized through an EMA qualification opinion [21, 22]. Re-analyses of AD trials reported that designs employing PROCOVA™ achieved equivalent statistical power with approximately 15–20% fewer patients; with the Bayesian extension PROCOVA+™, this reduction reached 50–66% in early-phase studies. These effect sizes are largely vendor-reported and derive from re-analyses conducted by the method developers; they have not yet been broadly replicated by independent groups, and their generalizability across disease areas, endpoints, and populations remains uncertain. Reported gains should therefore be read as favorable-case illustrations rather than as expected performance, and negative or null results remain under-represented in the currently available literature [6, 19]. These figures reflect outcomes of a specific condition in which the digital twin simultaneously leverages both covariate information and augmentation capacity — not an upper bound on power gains achievable through prognostic covariate control alone. Actual gains will vary substantially depending on the disease area, the prognostic predictability of the endpoint, and cohort homogeneity.

Sample size re-estimation (SSR): $n_{updated} = 2\sigma 2_{observed} (z_\alpha/2 + z_\beta)2 / \delta 2_{planned}$. A validated digital-twin prognostic model may improve interim variance estimation and thereby support more accurate sample-size re-estimation, provided that its calibration and transportability are maintained in the trial population; simultaneously, virtual patient augmentation may partially offset the operational impact of unexpected dropout, provided the added records are appropriately discounted rather than treated as independent observations. As a worked illustration, consider a two-arm trial powered at 80% (alpha = 0.05, two-sided) to detect a standardized effect of delta = 0.3, which requires approximately n = 175 per arm. If a validated digital-twin prognostic score explains a proportion R-squared = 0.20 of the outcome variance and enters the model as a PROCOVA-style covariate, the residual variance is scaled by (1—R-squared) and the required size falls to about n = 175 x (1—0.20) = 140 per arm—a gain that is realized only to the extent that R-squared holds in the actual trial population and vanishes if the score is poorly calibrated [21, 23].

Futility-Based Early Stopping

In group sequential designs, conditional power (CP) is the primary tool for futility assessment: $CP(t) = P(Z_final > z_\alpha \mid Z_t = z_t)$. Although CP < 20% is commonly used as the futility threshold, calibration to study-specific characteristics is recommended (1923). O'Brien-Fleming and Pocock boundary functions — foundational methods in adaptive trial design — provide pre-specified alpha-spending rules, while the Lan-DeMets framework controls the overall type-I error rate independently of the number and timing of interim analyses. Digital twin simulation can support prospective, design-stage exploration of trial performance, for example calibrating futility thresholds and adaptation rules under a range of plausible data-generating scenarios and model-misspecification conditions. Because a simulation reflects only the assumptions of the fitted model, it cannot determine the unobserved result the trial would have produced had enrollment continued, and retrospective simulations should not be interpreted as establishing whether a stopping decision was correct. Standard operating characteristics—type I error,

power, expected sample size, bias, and coverage—must be reported across these scenarios [1, 2].

Efficacy-Based Early Stopping

Early crossing of an efficacy boundary introduces a selection bias: because stopping occurs at a moment when the test statistic happens to be high, the observed effect size becomes an upward-biased estimator of the true population parameter. Median unbiased estimation (based on the conditional likelihood) can partially correct this bias, though the correctable portion is limited [23]. Regression to the mean is an equally critical constraint: the excessively large effect sizes observed early tend to approach the true population mean over time. The 'winner's curse' phenomenon — the failure of early positive trial results to be reproduced in subsequent larger trials — is well documented in the literature. Digital-twin simulation cannot recover the effect size that would have been observed had the trial continued, nor should it be presented as correcting the winner's curse or removing the bias introduced by efficacy-based early stopping. A digital twin is also unlikely to reliably identify late-emerging toxicities that are rare or poorly represented in its training data. Its more defensible role is prospective: exploring, at the design stage, how candidate stopping boundaries behave under a range of plausible scenarios [1, 2].

Table 1 Comparison of control arm strategies in clinical research

Control arm strategy	Data requirement	Strength	Key limitation
Randomized control	Concurrent randomized allocation	Unbiased causal inference by design	Cost, feasibility, and ethical limits in rare/pediatric settings
Historical control	Previously completed trials	No concurrent control required	Temporal and standard-of-care drift; confounding
External control	Real-world data / registries	Uses existing comparators; larger samples	Unmeasured confounding; comparability
Synthetic control (SCA)	Historical trial data + RWD	Fully model-generated comparator	Bounded by source data; model-dependent validity
Digital twin / ICA	Multimodal individual patient data	Patient-specific predicted control; improves precision when calibrated	Model-based prediction, not randomization; transportability-dependent

Comparison of control arm strategies used in clinical trials. Traditional randomized control arms are contrasted with historical, external, synthetic, and digital twin–based control approaches

External Control Arms and Intelligent Control Arms (ICA)

The ethical defensibility of a randomized control arm is increasingly questioned in rare diseases, pediatric populations, and areas of high unmet medical need. In the ECA approach, propensity score matching (PSM) and inverse probability of treatment weighting (IPTW) attempt to establish comparability: $e(X) = P(T=1 \mid X)$; however, these methods remain vulnerable to unmeasured confounders [24]. In the Intelligent Control Arm (ICA) design, the control arm is composed of each active-arm patient's digital twin — that is, the predicted placebo outcome that patient would have exhibited had they been assigned to the control group. Since each patient is compared against their own digital control, patient-specific prognostic predictions may explain part of the baseline outcome heterogeneity and improve precision when the model is well calibrated (Table 1). A digital control is a model-generated prediction of an unobserved potential outcome rather than a randomized comparator; its validity depends on whether the model captures the relevant prognostic variables and treatment-effect modifiers and on whether it is transportable from its training population to the treated population, so it does not recreate randomization or, by itself, guarantee unbiased causal estimation [1, 6]. The ICA design combines ethical and statistical advantages: all patients have the opportunity to receive active treatment, while patient-specific prognostic prediction can improve precision and reduce reliance on between-group matching, without recreating randomization or guaranteeing unbiased causal estimation [1].

Hybrid Designs: Augmented and Synthetic Control

In the Augmented Control (AC) design, digital twins are added to real control patients to increase statistical power; in the Synthetic Control Arm (SCA) design, the control arm is constructed entirely from historical trial databases and RWD (Table 1, Fig. 3). When historical or external data are balanced against the trial population, an absolute SMD below 0.1 is commonly used as a diagnostic criterion. Within the power prior framework, the weight coefficient α may be prespecified or estimated according to the selected borrowing framework: $\alpha = 1$ denotes full weighting, $\alpha = 0$ denotes exclusion of historical data [23]. Federated learning offers the opportunity to train AI models across institutions without directly sharing trial data; However, between-study data heterogeneity continues to constitute a however, between-study data heterogeneity continues to constitute a significant practical constraint [1, 2].

Comparison of traditional randomized controlled trials, external or synthetic control approaches, and digital

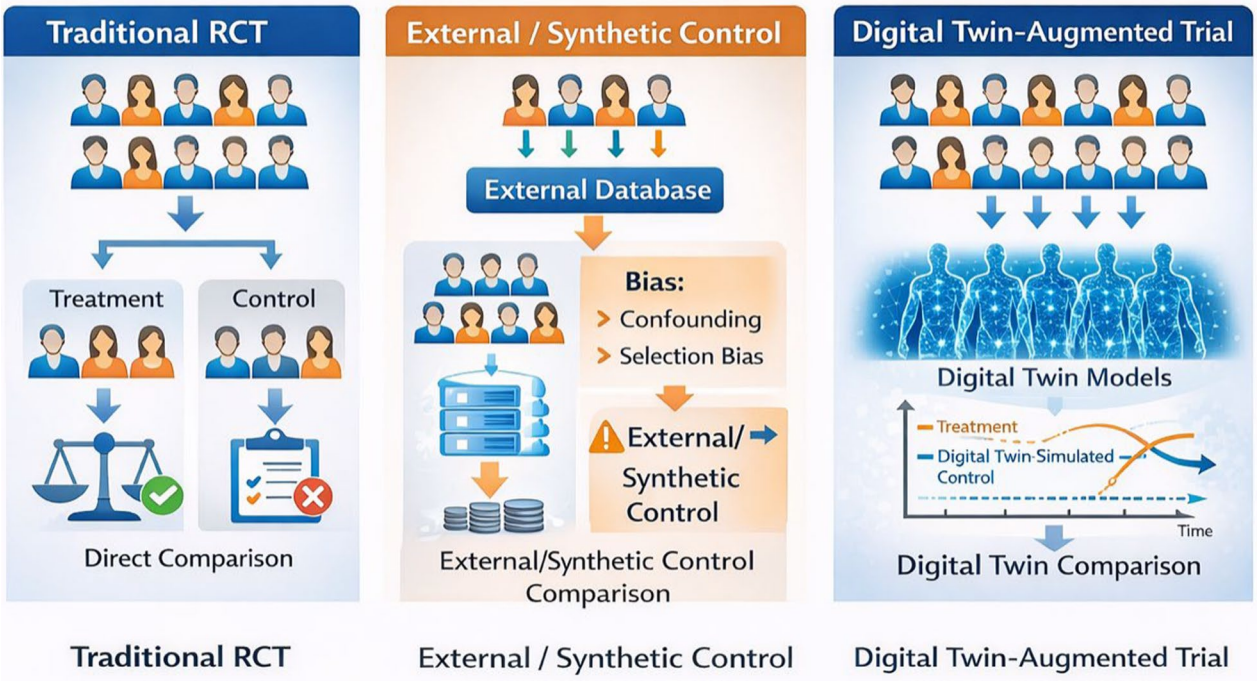


Fig. 3 Digital twin–augmented clinical trial design

twin–based simulations used to generate individualized model-predicted control trajectories.

Pre-Trial In Silico Design and Protocol Optimization

The computer-aided trial design taxonomy encompasses three components: trial semantic parsing (automatic extraction of PICO elements from protocol free-text via NLP), outcome prediction (trial success/failure forecasting using multi-modal real-world data; Liu et al. reported AUROC 0.75 in oncology trials), and protocol evaluation (computing the marginal contribution of individual eligibility criteria to cohort size and hazard ratio via Shapley scores; clustering methods can identify superior criterion combinations) [2, 25].

The statistical layers of pre-trial simulation encompass: scenario-based power calculations under varying assumptions of effect size, dropout rate, and treatment adherence profile; interim analysis timing optimized via O'Brien-Fleming or Lan-DeMets spending functions; determination of the effect of SSR rules on conditional power through Monte Carlo simulation; and simulation of subgroup size and treatment-biomarker interaction in biomarker-based enrichment strategies [23, 25].

Indirect Treatment Comparisons and Digital Twin–Based Simulation

When direct head-to-head RCTs are unavailable, researchers rely on indirect treatment comparison (ITC) methods. The simplest approach, the Bucher indirect comparison, estimates treatment effects through a shared comparator arm and assumes that study populations are sufficiently comparable—an assumption that may be violated in practice [26]. Network meta-analysis (NMA) extends indirect comparisons by integrating results from a network of studies, allowing simultaneous comparison of multiple treatments, but relies on assumptions of population similarity and consistency between direct and indirect evidence [27].

Population-adjusted indirect comparison (PA-ITC) techniques such as matching-adjusted indirect comparison (MAIC) and simulated treatment comparison (STC) aim to correct for baseline differences between study populations. MAIC uses weighting to align individual patient data from one study with the aggregate characteristics of another: $\theta_MAIC = \Sigma(w_i \times Y_i) / \Sigma(w_i)$; the key limitation is reduction in effective sample size (ESS). STC applies regression models to predict outcomes in the target population [28, 29]. Although these approaches improve comparability, they remain limited because they can adjust only for measured covariates. IPD-NMA can decompose both within-study and between-study heterogeneity, but data access difficulties constrain its practical use [30–32].

Digital twin–based simulation has recently been proposed as a complementary strategy for comparative effectiveness evaluation. Patient-specific computational models constructed using multimodal clinical data generate patient-specific model-predicted control trajectories under standardized baseline conditions, enabling in silico comparisons between treatments evaluated in separate studies and avoiding the weighting-induced ESS reduction that affects MAIC, while remaining dependent on model validity and the information contained in the training data [1]. Importantly, digital twin approaches should not be viewed as replacements for established statistical methods. Combining digital twin simulations with established PA-ITC methods may offer a more robust framework for comparative effectiveness evaluation, particularly in settings where direct randomized comparisons are unavailable (Table 2) [32].

Discussion

The integration of digital twin technology into clinical research methodology must be evaluated along three axes: technical maturity, regulatory acceptance, and statistical methodology. Vidovszky et al. argue that the highly regulated structure of the clinical trials domain provides the ideal environment for the safe implementation of this technology; regulatory acceptance achieved in clinical trials will pave the way for broader application of AI-generated digital twins in clinical practice [1].

Technical constraints can be organized under four headings: [1] each patient's unique health profile, shaped by genetic, environmental, and lifestyle factors, differs fundamentally from standardized mechanical systems; the standard of care and access to new therapies are continuously evolving [1, 2] EHRs offer rich data but are constrained by format inconsistencies and poorly labeled disease definitions; data drift and changes in standard of care can erode model reliability over time [1, 3, 7] models must handle all variable types, operate in the presence of missing data, and capture complex inter-variable interactions [2, 4] patient data are protected by legal frameworks ensuring both privacy and consent for use in clinical research [33].

Table 2 Methods for indirect treatment comparison in clinical research

Method	Data requirement	Strength key	Limitation
Bucher indirect comparison	Aggregate trial data	Simple and widely used	Sensitive to population differences
Network meta-analysis [NMA)	Multiple RCTs	Enables comparison of many treatments	Requires consistency across studies
MAIC	IPD + aggregate data	Adjusts for baseline population differences	Reduced effective sample size
STC	Individual patient data	Flexible modeling of treatment effects	Model specification sensitive
Digital twin simulation	Multimodal patient data	Patient-specific model-predicted control trajectories; avoids MAIC's weighting-induced ESS reduction	Dependent on model validity

Algorithmic bias constitutes a particularly important constraint. Models that insufficiently represent populations historically underrepresented in clinical trials — specific ethnic groups, older adults, women, and low-income countries — may lose reliability for these groups [1, 8]. From a complex systems perspective, biological variability, multi-scale dynamics, and epistemic uncertainty in model structure further compound these challenges: the interaction between data sparsity and model complexity represents one of the central open problems in translating digital twin research into routine clinical practice [34]. Approaches such as PROCOVA™ and PROCOVA+™ provide standardized solutions for integrating digital twins into statistical inference; however, inference in sparse data scenarios (Phase I, rare diseases) and inter-institutional heterogeneity in federated learning applications remain open methodological challenges [2, 21].

This perspective differs from previous reviews in both emphasis and organization. Whereas earlier reviews have been oriented primarily toward the underlying technology, the present article organizes the literature around nine concrete statistical and methodological challenges of clinical trials. On this basis it proposes a conceptual integration of digital-twin simulation with population-adjusted indirect treatment comparison methods (MAIC, STC, and IPD-NMA), which we regard as the manuscript's principal original contribution (Table 3).

Despite these potential advantages, several important limitations remain. Digital twin models depend heavily on the availability of high-quality and representative datasets, and biases in training data may affect predictive performance across diverse patient populations. In addition, regulatory frameworks for integrating AI-generated simulations into clinical trial design are still evolving [35]. For these reasons, digital twin approaches should currently be viewed as complementary methodological tools rather than replacements for traditional randomized trials. Continued methodological research, prospective validation studies, and collaboration with regulatory agencies will be essential to define their future role in clinical research.

Table 3 Clinical trial challenges, current approaches, and digital twin contributions

Digital twin contribution	Digital twin contribution	Digital twin contribution
Patient dropout (MNAR)	Reference-based imputation, pattern-mixture models, selection models, and tipping-point sensitivity analyses	Model-based sensitivity analysis generating plausible trajectories; input sensitivity
Partial participation	ITT, censoring	Prognostic covariate adjustment; related Bayesian extension (e.g., PROCOVA+™)
Inadequate sample size	Prospective sample-size calculation, sample-size re-estimation, recruitment extension, and adaptive design	Virtual augmentation, prognostic covariate control, SSR
Futility early stopping	Conditional power (CP)	Prospective design-stage CP simulation; protocol evaluation
Efficacy early stopping	Median unbiased estimation	Prospective boundary calibration; limited for rare/late toxicity
Control arm ethics	Historical control, PSM, IPTW	ICA — patient-specific model-predicted control trajectory
Hybrid designs	Augmented/Synthetic control	Virtual cohort + power prior; federated learning
Trial design optimization	Conventional power analysis	Computer-aided design; Shapley protocol evaluation
Indirect comparison	NMA, Bucher, MAIC, STC, IPD-NMA	PA-ITC complement: avoids weighting-induced ESS reduction; depends on model validity

Conclusion

The nine methodological challenges addressed in this perspective article concretely frame the pressures bearing on the statistical validity and clinical reliability of clinical research. Digital twin technology has undergone a fundamental transformation along a developmental arc that extends from Grieves' 2002 industrial concept through NASA's aerospace applications to multi-omics medical models [1, 3].

As recent systematic frameworks demonstrate, the three foundational pillars of in silico clinical trials—clinical simulation, individualized predictive modeling, and computer-aided trial design—constitute a coherent methodological ecosystem. The concrete applications documented by Mollica et al. in the oncology context, the regulatorily endorsed PROCOVA™ framework presented by Unlearn.ai, and the rapidly growing body of healthcare digital twin literature—with publications increasing each year from 2017 to 2023—demonstrate that digital twins are beginning to generate practical evidence beyond their theoretical potential [2, 4, 6, 8].

Four prerequisites will be decisive for this technology to reach its full methodological potential in clinical research: establishment of consistent validation and explainability standards (including SHAP, LIME, and independent cohort validation); development of methodological frameworks for the detection and mitigation of algorithmic bias; clarification of regulatory acceptance criteria; and construction of comprehensive data infrastructure to address representativeness gaps. Interdisciplinary collaboration spanning medicine, engineering, biostatistics, ethics, and law is an indispensable condition for achieving these goals [1, 8, 33]. Rather than replacing randomized clinical trials, digital twins are likely to become complementary methodological tools whose ultimate value will depend on rigorous validation, transparent reporting, and regulatory acceptance.

Author contributions Fatma Nur Cicin: Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Data Curation, Writing—Original Draft Preparation, Writing—Review & Editing, Visualization, Project Administration.

Funding This study received no specific funding.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethical approval Not applicable.

References

1. Vidovszky AA, Fisher CK, Loukianov AD, Walsh JR, Unlearn Research Team. Increasing acceptance of AI-generated digital twins through clinical trial applications. Clin Transl Sci. 2024;17:e13897. https://doi.org/10.1111/cts.13897.
2. Akbarialiabad H, Pasdar A, Murrell DF, et al. Enhancing randomized clinical trials with digital twins. npj Syst Biol Appl. 2025;11:110. https://doi.org/10.1038/s41540-025-00592-0.
3. Grieves M, Vickers J. Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In: Kahlen J, Flumerfelt S, Alves A, editors. Transdisciplinary perspectives on complex systems. Springer; 2017. p. 85–113. https://doi.org/10.1007/978-3-319-38756-7_4.
4. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. npj Digit Med. 2024;7:77. https://doi.org/10.1038/s41746-024-01073-0.
5. Sadée C, Testa S, Barba T, et al. Medical digital twins: enabling precision medicine and medical artificial intelligence. Lancet Digit Health. 2025;7(7):100864. https://doi.org/10.1016/j.landig.2025.02.004.
6. Unlearn.ai. (2022). Applications of digital twins in clinical trials for Alzheimer's disease: How digital twins make it possible to design and run more efficient clinical trials with well-defined statistical properties [White paper]. Unlearn.ai.

7. Viceconti M, Henney A, Morley-Fletcher E. In silico clinical trials: how computer simulation will transform the biomedical industry. Int J Clin Trials. 2016;3(2):37–46. https://doi.org/10.18203/2349-3259.ijct20161408.
8. Mollica L, Leli C, Sottotetti F, Quaglini S, Locati LD, Marceglia S. Digital twins: a new paradigm in oncology in the era of big data. ESMO Real World Data Digit Oncol. 2024;5:100056. https://doi.org/10.1016/j.esmorw.2024.100056.
9. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
10. Tardini E, Zhang X, Canahuate G, et al. Optimal treatment selection in sequential systemic and locoregional therapy of oropharyngeal squamous carcinomas: deep Q-learning with a patient-physician digital twin dyad. J Med Internet Res. 2022;24(4):e29455. https://doi.org/10.2196/29455.
11. Chaudhuri A, Pash G, Hormuth DA 2nd, et al. Predictive digital twin for optimizing patient-specific radiotherapy regimens under uncertainty in high-grade gliomas. Front Artif Intell. 2023;6:1222612. https://doi.org/10.3389/frai.2023.1222612.
12. Wu C, Lorenzo G, Hormuth DA 2nd, et al. Integrating mechanism-based modeling with biomedical imaging to build practical digital twins for clinical oncology. Biophysics Reviews. 2022;3(2):021304. https://doi.org/10.1063/5.0086789.
13. Voigt I, Inojosa H, Dillenseger A, Haase R, Akgün K, Ziemssen T. Digital twins for multiple sclerosis. Front Immunol. 2021;12:669811. https://doi.org/10.3389/fimmu.2021.669811.
14. Laubenbacher R, Mehrad B, Shmulevich I, Trayanova N. Digital twins in medicine. Nature Computational Science. 2024;4(3):184–91. https://doi.org/10.1038/s43588-024-00607-6.
15. Fisher CK, Smith AM, Walsh JR, et al. Machine learning for comprehensive forecasting of Alzheimer's disease progression. Sci Rep. 2019;9(1):13622. https://doi.org/10.1038/s41598-019-49656-2.
16. Medcalf E, Turner RM, Espinoza D, He V, Bell KJL. Addressing missing outcome data in randomised controlled trials: a methodological scoping review. Contemp Clin Trials. 2024;143:107602. https://doi.org/10.1016/j.cct.2024.107602.
17. Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. Wiley; 2019. https://doi.org/10.1002/9781119482260.
18. ICH E9(R1). (2019). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. International Council for Harmonisation.
19. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16:29. https://doi.org/10.1186/s12916-018-1017-7.
20. Pammi M, Burdjalov V, Lim J, Morales O. Digital twins, synthetic patient data, and in-silico trials: can they empower paediatric clinical trials? Lancet Digit Health. 2025;7(5):100851. https://doi.org/10.1016/j.landig.2025.01.007.
21. Schuler A, Walsh D, Hall D, Walsh J, Fisher C. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. Int J Biostat. 2022;18(2):329–56. https://doi.org/10.1515/ijb-2021-0072.
22. European Medicines Agency. (2022). Qualification opinion for PROCOVA™ prognostic covariate adjustment. EMA/CHMP/SAWP/784219/2022.
23. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. Stat Med. 2016;35(3):325–47. https://doi.org/10.1002/sim.6472.
24. Food and Drug Administration. (2023). Considerations for the design, conduct, and analysis of observational studies using real-world data: Draft guidance for industry. U.S. Department of Health and Human Services.
25. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and artificial intelligence. Nature. 2021;592(7855):629–33. https://doi.org/10.1038/s41586-021-03430-5.
26. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network meta-analysis for decision-making. John Wiley & Sons; 2018. https://doi.org/10.1002/9781118951651.
27. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods. 2012;3(2):80–97. https://doi.org/10.1002/jrsm.1037.
28. Ciani O, Amorim DL, Capri S, et al. Reflecting on a decade since the introduction of matching-adjusted indirect comparison: a review of the methodological developments. Value Health. 2021;24(11):1680–5. https://doi.org/10.1016/j.jval.2021.08.007.
29. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. Med Decis Making. 2018;38(2):200–11. https://doi.org/10.1177/0272989X17725740.
30. Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. J R Stat Soc A. 2020;183(3):1189–210. https://doi.org/10.1111/rssa.12579.
31. NICE Decision Support Unit. (2016). NICE DSU technical support document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. National Institute for Health and Care Excellence.
32. Remiro-Azócar A, Heath A, Baio G. Methods for population adjustment with limited access to individual patient data: a review and simulation study. Res Synth Methods. 2021;12(6):750–75. https://doi.org/10.1002/jrsm.1511.
33. Huang PH, Kim KH, Schermer M. Ethical issues of digital twins for personalized health care service: preliminary mapping study. J Med Internet Res. 2022;24(1):e33081. https://doi.org/10.2196/33081.
34. De Domenico M, Allegri L, Caldarelli G, et al. Challenges and opportunities for digital twins in precision medicine from a complex systems perspective. npj Digit Med. 2025;8:37. https://doi.org/10.1038/s41746-024-01402-3.
35. European Medicines Agency. (2023). Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. EMA/83588/2023.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.