CPT: Pharmacometrics & Systems Pharmacology

REVIEW OPEN ACCESS

Synthetic Data in Healthcare and Drug Development: Definitions, Regulatory Frameworks, Issues

Giuseppe Pasculli[1] | Marco Virgolin[2] | Puja Myles[3] | Anna Vidovszky[4] | Charles Fisher[4] | Elisabetta Biasin[5] | Miranda Mourby[6] | Francesco Pappalardo[7] | Saverio D'Amico[8,9] | Mario Torchia[1] | Alexander Chebykin[2] | Vincenzo Carbone[1] | Luca Emili[1] | Daniel Roeshammar[1]

[1]InSilicoTrials Technologies S.p.A., Trieste, Italy | [2]InSilicoTrials Technologies B.V., s-Hertogenbosch, the Netherlands | [3]Medicines and Healthcare products Regulatory Agency, London, UK | [4]Unlearn.AI, San Francisco, California, USA | [5]Centre for IT & IP Law (CiTiP), KU Leuven, Leuven, Belgium | [6]Centre for Health, Law, and Emerging Technologies (HeLEX), Faculty of Law, University of Oxford, Oxford, UK | [7]Department of Drug and Health Sciences, University of Catania, Catania, Italy | [8]Humanitas Clinical and Research Center-IRCCS, Milan, Italy | [9]Train s.r.l., Milan, Italy

Correspondence: Giuseppe Pasculli (rmhaggp@ucl.ac.uk)

Received: 31 January 2025 | **Revised:** 1 March 2025 | **Accepted:** 10 March 2025

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. All aspects of this study, including design, data collection, analysis, and manuscript preparation, were conducted by the authors independently and without financial support.

Keywords: drug development | external control arms | generative AI | medical devices | real-world data | regulatory framework | synthetic data

ABSTRACT

With the recent and evolving regulatory frameworks regarding the usage of Artificial Intelligence (AI) in both drug and medical device development, the differentiation between data derived from observed ('true' or 'real') sources and artificial data obtained using process-driven and/or (data-driven) algorithmic processes is emerging as a critical consideration in clinical research and regulatory discourse. We conducted a critical literature review that revealed evidence of the current ambivalent usage of the term "synthetic" (along with derivative terms) to refer to "true/observed" data in the context of clinical trials and AI-generated data (or "artificial" data). This paper, stemming from a critical evaluation of different perspectives captured from the scientific literature and recent regulatory endeavors, seeks to elucidate this distinction, exploring their respective utilities, regulatory stances, and upcoming needs, as well as the potential for both data types in advancing medical science and therapeutic development.

1 | Introduction

In the face of rapidly growing data challenges in the global healthcare sector, such as privacy concerns, confidentiality, data fragmentation, validity questions, interoperability, and generalizability issues, synthetic data are stepping forward as a potential source of innovation. The European Health Data Space (EHDS) proposal (entered into force in March 2025 and published in the European Official Journal) by the European Commission aims to establish a unified data market, leveraging health data for care delivery, research, and policy development [1]. Synthetic data may be a key element in this initiative, as they can facilitate scientific

Abbreviations: AI, Artificial Intelligence; CBER, Center for Biologics Evaluation and Research within FDA; CDER, Center for Drug Evaluation and Research within FDA; CDRH, Center for Devices and Radiological Health within FDA; CM&S, Computer Modeling and Simulation; CPRD, Clinical Practice Research Datalink; DGA, Data Governance Act; DMs, Diffusion Models; ECA, External Control Arm; EHDS, European Health Data Space; EHR, Electronic Health Record; EMA, European Medicines Agency; ER, Exposure-Response model; FDA, US Food and Drug Administration; GANs, Generative Adversarial Networks; GDPR, General Data Protection Regulation; ICH, International Conference on Harmonization; ISPE, International Society for Pharmacoepidemiology; IVDR, In Vitro Diagnostic Medical Devices Regulation; MHRA, Medicines and Healthcare products Regulatory Agency; ML, Machine Learning; ODE, Ordinary Differential Equation; PBPK, Physiologically based Pharmacokinetic model; PD, Pharmacodynamic; PK, Pharmacokinetic; popPK, Population Pharmacokinetic model; popPKPD, Population Pharmacokinetic/Pharmacodynamic model; QSP, Quantitative Systems Pharmacology; RCTs, Randomized Controlled Trials; RWD, Real-World Data; SCAs, Synthetic Control Arms; US, United States; VAEs, Variational Autoencoders; VCAs, Virtual Control Arms.

Disclaimer: Puja Myles's contributions reflect her own views and not those of the MHRA.

advancement without compromising data privacy if generated carefully. However, the absence of a universally accepted definition for synthetic data complicates regulatory efforts, making it crucial to establish clear terminology in the rapidly evolving landscape of data usage and privacy. As an example, the United States (US) Census Bureau presents synthetic data in computer programming as entirely simulated data constructs for testing, free from real-world constraints. At the same time, they recognize that in statistics, synthetic data often depict the amalgamation of multiple sources to produce detailed estimates [2].

We define synthetic data in line with the glossary of US Food and Drug Administration (FDA) on Digital Health and Artificial Intelligence [3] as, quoting:

> Data that have been created artificially (e.g., through statistical modeling, computer simulation) so that new values and/or data elements are generated. Generally, synthetic data are intended to represent the structure, properties, and relationships seen in actual patient data, except that they do not contain any real or specific information about individuals.

Importantly, we define observed (or 'true') data as data that are obtained by direct measurement of collection from real-world events (hence including Randomized Controlled Trials (RCT) and Real-World Data (RWD)), and that are typically used as input to produce synthetic data [4].

In line with a recent discussion on the topic by Selvarajoo & Maurer-Stroh [5], for a clearer distinction, we also pose that synthetic data can be broadly categorized into process-driven and data-driven approaches:

- Process-driven synthetic data are generated using computational or mechanistic models based on biological or clinical processes and has been an established and regulatory-accepted paradigm for decades [6-8]. These models typically use known mathematical equations (e.g., ordinary differential equations (ODEs)), such as pharmacokinetic (PK) and pharmacodynamic (PD) models and agent-based simulations [9,10]. The models are first developed to explain an observed behavior and are then subsequently used to generate simulated or synthetic data using the same model for different conditions or situations [11].
- Data-driven synthetic data rely on statistical modeling and machine learning (ML) techniques, including sequential ensembles of decision trees, Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs), that have been trained on actual ("observed") data and create synthetic datasets that preserve population-level statistical distributions.

Nevertheless, there is still no generally accepted terminology when it comes to "synthetic" data, particularly when referring to data constituting external control arms (ECAs) in the context of clinical trials and drug development. According to the FDA draft guidance in 2023 [12], the ECA is defined as "..a group of people, treated or untreated, from an earlier time (historical control), ..., or during the same period (concurrent control) but in another setting." It's worth noting that this draft guidance does not currently address whether an ECA can be composed of artificial or synthetic data; indeed, "synthetic control arms (SCAs)" or "virtual control arms (VCAs)" have also been used, we might argue "incorrectly" in light of the above definitions provided, as synonyms of ECAs [13,14]. As an example, and as will be showcased in the upcoming sections of this manuscript, the ECA approach has been indicated with mixed nomenclatures when harnessing observed data for external controls from sources such as electronic health records (EHRs), administrative claims, patient-generated information, disease registries, and prior clinical trial data [15], all with the final aim to offer an alternative to internal control groups [16].

1.1 | RWD For ECAs: An Established Paradigm?

Despite the option to seek a (randomized) concurrent control still representing the gold standard in drug development, ECAs obtained from RCT or RWD sources have gained substantial traction particularly in contexts of providing supportive evidence (rather than confirmatory), with numerous drugs (particularly for rare diseases and unmet medical needs) approved to market through this approach [17], exemplifying the utility of such (observed) data from actual patient experiences. These ECAs serve as a critical component in observational studies and for comparative analyses where RCTs may be infeasible or unethical [18]. Indeed, from a regulatory standpoint, the FDA's publication of guidelines specific to the use of external data sources in constructing ECAs for drug development constituted another step towards the acceptance of RWD and experimental cohorts role in this approach [12]. As for the European context, the study by Wang et al. [19] provided a focused review, indicating that 18 European Medicines Agency (EMA)-approved oncology drugs incorporated 24 ECAs (defined within that manuscript as "data derived outside the concurrent clinical trial") obtained from external data sources from 2016 to 2021. Despite this progress, the authors also identified critical hurdles that such data face in the context of EMA evaluations: about one-third of the ECAs were not considered supportive by the EMA, often due to issues related to lack of patient population heterogeneity and gaps in outcome assessments within the external data sources.

1.2 | AI-Generated Data: New Frontier With New Complexities

In the context of data-driven (hence not process-driven) generation processes [5], generated synthetic data (also referenced as "artificial" [20] or "simulated" data [21,22]) have been in use for many years, albeit to a lesser extent than today. An example is imputation, i.e., filling in missing values: Artificial Intelligence (AI) models allow users to go beyond simpler heuristics such as using the mean (for numerical) or mode (for categorical) value by learning how other features influence the value to be filled in. Dong et al. provide an example of this as an application in healthcare [23].

However, it is only in the last decade that methodological advances in AI (e.g., the attention mechanism [24]) and sufficient computing power [25] have made it possible to reliably generate high-fidelity synthetic values at the level of *entire datasets*. Another important enabling factor is *self-supervised pre-training*: a family of techniques to train an AI model to be generally useful on a variety

of tasks by utilizing large amounts of data, without the need for human annotations [26,27]. When it comes to clinical trials, pretraining may enable synthetic data generation for ECAs, as data from control arms of different trials may be combined in a sufficiently large dataset to feed these algorithms [16].

Modern, data-driven, generative AI models include GANs, VAEs, Diffusion Models (DMs), and Transformers [24,28,29,30]. These models operate in two phases. Firstly, a model is trained using observed data: model parameters are adjusted such that the synthetic data produced by the model is similar to the original data. After the training, the model parameters are fixed, with the model now ready to generate synthetic data with statistical properties that are quasi-identical to those of the original observed data source [31] without the generated data being directly linked to any particular individual present in the originating data [32]. In this second phase (inference), the AI model can be used to generate, or "sample," synthetic data at will. For a more detailed review on the topic, we refer to [33].

With the advent of generative AI, the terminology of what constitutes "synthetic" data has been extended further, with some authors, we argue righteously in light of the rationale and definitions above, referring to "synthetic" as data created via generative AI techniques and/or process-driven methods [5,31], [34-37]. Therefore, it has become important to distinguish between what was previously called "synthetic" (i.e., patient data possibly collected from myriad sources) and data that are generated (i.e., artificial data) by data- or process-driven methods [5] to avoid confusion about the type of data being actually used [31].

1.3 | Study Rationale

This paper endeavors to map, by means of critical literature review, the usage of the term "synthetic" with reference to data, delineating its provenance from observed ('true')-derived constructs to generated data. We also reference *in silico* trial approaches that may involve the use of simulated data or virtual patient cohorts, framing these within more established process-driven paradigms of synthetic data [5]. Finally, we analyze the evolving landscape of terminology in synthetic data research and propose a framework to mitigate ambiguity in its interpretation and application.

2 | Methodology

This review employs a critical narrative approach to explore the understanding of terminology for different data sources in healthcare and drug development settings. Unlike systematic reviews that focus on answering specific, narrow questions through predefined methods, a critical narrative review allows for a broader examination of diverse studies, providing interpretation and critique across a wider scope of literature [38,39].

2.1 | Search Strategy

The literature search was conducted from 1986 to 2025 using the PubMed database to identify relevant studies. The following search query was employed to retrieve articles:

$$(("\text{control arm}" \text{ AND } ("\text{synthetic}" \text{ OR } "\text{external}" \text{ OR } "\text{virtual}" \text{ OR } "\text{simulated}" \text{ OR } "\text{in silico}") \text{ AND } "\text{data}"))$$

This query was designed to capture studies discussing control arms that utilize either synthetic or external data and other possible derivatives thereof. By specifying "control arm" and "data," the search focused on relevant research involving these data types. The use of "synthetic" OR "external" OR "virtual" etc., broadened the scope to include various process-driven methodologies spanning from more established contexts [40], ensuring a review of how these data types are referred to in medical and scientific research.

The search returned a total of 208 results, which were then screened in terms of content to make sure that the literature contained relevant information for the rationale of the manuscript. The final selected results were then summarized in Table S1 of this manuscript, accounting for $n = 91$ instances.

3 | Discussion

3.1 | The Multiplicity of "Synthetic Data"

The literature research revealed a bifurcation in the use of this term. On one side, synthetic data, in line with the data-driven intuition proposed in Selvarajoo & Maurer-Stroh [5], are referenced as generative AI outputs—artificial constructs devised through advanced computational models, such as GANs. These artificially produced datasets, also sometimes labeled as "false" [41] or "fake" data in different contexts [42-44], serve several purposes, primarily in exploratory and modeling capacities to simulate scenarios, patterns, or outcomes that may not be feasible or ethical to generate through traditional clinical trials. Pioneering studies by authors like Azizi et al., El Kabbaj et al., Fisher et al., and D'Amico et al. have contributed to this growing body of knowledge, pushing for the usage and recognition of this artificial data (defined as *synthetic data* in their works) as a possible proxy for observed ('real') data in different therapeutic areas [20,45,46,47]. It is noteworthy that other recent works have also adopted the same notation [34,35,36,37,41,48] and other authors like Alloza et al. have also sought to establish the role of artificial data (still referred to as *synthetic data*) in shaping regulatory decision-making processes [31].

Interestingly, a narrative review by Gonzales, Guruswamy, and Smith also pointed out that the term "synthetic data" has been widely used to characterize datasets in various synthesized forms and levels [49]. They also describe three broad categories of synthetic data, specifically: (i) *Fully Synthetic Data*: Data that are completely artificial and do not contain any real data; (ii) *Partially Synthetic Data*: Describes datasets where only certain sensitive variables are replaced with synthetic counterparts, hence maintaining some level of real data; and (iii) *Hybrid Synthetic Data*:

Data created by combining both real and synthetic data [49]. While partially synthetic data modify only selected attributes within real datasets (hence risk for reidentification is still present), hybrid synthetic data blend entire synthetic records with real records, offering strong privacy protection while maintaining high utility compared to the first two categories.

On the other hand, based on our findings, the term RWD was generally understood as referring to authentic ("observed") patient data, but has been also used in combination with the term "synthetic", particularly in the context of clinical trials (e.g., single-arm trials). Indeed, Boyne et al. and Van Le et al. used the term "Real-World Synthetic Control Arm" to describe RWD (obtained respectively from a national cancer registry and various clinical sites/research database sources) when used to construct comparative analyses in the absence of traditional randomized control arms [50,51]. Similarly, Popat et al., along with Banerjee et al., Burcu et al., O'Haire et al., Neehal et al., Thorlund et al., Yoshino et al., and Zhu et al., opted for the designation "Synthetic Control Arm" or "Synthetic Controls" [52], eschewing references to observed data with such nomenclature [53-60]. Interestingly, the work by Burcu et al. [53], also endorsed by the International Society for Pharmacoepidemiology (ISPE), posed the perspective that "External control arms are also called 'synthetic' control arms as they are not part of the original concurrent patient sample". While this definition provides an endorsed framework within the context of external controls, we believe it does not fully encompass the broader scope introduced by recent AI-driven methodologies for generating synthetic data. As these innovations continue to redefine the landscape of data generation in clinical research, a shared discussion with scientific societies becomes increasingly relevant to refine definitions and ensure alignment with emerging technological and regulatory perspectives. Meanwhile, in their work Uemura et al., used the term "External Synthetic Control" to refer to RWD whereas Serrano et al. pose that "External control arms include patient-level real-world data, prospective cohorts or registries, and *synthetic control arms* elaborated from pooled or individual clinical trial data", adding another layer of complexity (and, potentially, confusion) to the matter [13,61]. Menefee et al. and Walker et al. [62,63] define data from previously conducted randomized trials (hence true/observed [64]) as "Synthetic Control Arms", whereas Davi et al. define SCAs an "external control constructed from patient-level data from previous clinical trials to match the baseline characteristics of the patients in an investigational group and can augment a single-arm trial" [65,66].

Another perspective, possibly in contrast with the concept of "*hybrid*" previously elucidated by Gonzales, Guruswamy, and Smith [49], was noted in Li et al. [67] defining the mixture of RWD and clinical trial data as a "hybrid control arm", similarly to Tan et al., Sengupta et al., Zou et al., and Neehal et al. [58,67,68,69,70], but clashing with the definition by Kurki et al. [15], wherein the combination of RWD with RCT data is simply defined as an ECA.

In contrast to the previous mixed scenario of definitions, similarly to the above mentioned work by Kurki et al. [15], the majority of papers included in this critical review simply referred to a observed (RWD or RCTs) data-composed control arm as an "ECA", hence possibly pointing toward an emerging consensus over the usage of such terminology [71-80]; please see Table S1 for a complete list of references.

Keeping in mind once again the proposed distinction between synthetic data as generated by a data-driven vs. a process-driven methodology [5], the discourse on data in clinical research acquires an additional layer of complexity with the introduction of the term "virtual controls" as in Switchenko et al. and Strayhorn [14,81]. "Virtual controls," as per Strayhorn and Switchenko et al., involve the use of actual observed outcome data from untreated individuals, coupled with statistical techniques to create counterfactual scenarios, thus offering a comparative baseline without the ethical concerns of withholding treatment. In another description, the "virtual control arms" nomenclature was utilized to refer to a deep learning (hence data-driven) algorithm trained on data from historical control patients and able to generate a likely outcome in the form of biomarker status or a clinical endpoint [82]. In another work, Chen et al. defined virtual control arm as generated by bootstrapping observed data with replacement, whereas Nicholson et al. adopted such nomenclature when referring to data generated by a machine learning prediction algorithm [83,84], hence referring to data-driven methods for synthetic data generation [5]. Differing from the previous work, and pointing towards process-driven methods for synthetic data generation [5], Folse et al. and Visentin et al. exploited the terminology of *virtual patients* as data generated by an ODEs model representing the physiological and disease pathways in cardiovascular events [85] and model generated data, respectively [86], whereas Dutta et al. adopted the "simulated historical control" when referring to data obtained via bootstrap of a Phase III trial data [87]. In another work, the ECA was referred to in a more generic way as "patients collected from data sources external to the single-arm trial," with synthetic data referred to as data obtained by means of "synthetic simulations" [88]. Ultimately, Suissa utilized "simulated data" to refer to data generated via exponential, and survival outcome distributions, whereas McMahon et al. adopted the terminology of "simulated study arm" to refer to study arms composed of simulated patients generated by a state-transition model analyzed as patient-level Monte-Carlo simulation [89,90].

To conclude, in their work also co-authored by an FDA member, Seeger et al., in contrast with ISPE's endorsed perspective by Burcu et al. [53], addressing that "synthetic controls is sometimes used interchangeably with external control groups", articulated that the dual use of "synthetic controls" when referring to observed ('true') data can lead to ambiguity, especially with the implication that the data might be "partially fabricated" [16]. To mitigate this confusion and ensure clarity ("*Due to the potential for confusion across these uses…*"), the term "external control group" was preferred by Seeger et al. when describing observed data that serve as a benchmark or point of reference in observational studies or clinical trials. The same nomenclature of "external control groups" or "arm" was preferred in other works [19,91,92] (see Table S1 for a complete reference list), as well as in a recent systematic review on the use of ECAs in immune-mediated inflammatory diseases [93]. Acknowledging the lexical overlap between "external control," "historical control," and "synthetic control" observed in the literature, Wang et al., in line with the International Conference on Harmonization (ICH) definition (2000), align with the view that "external control" should be

the term of choice for controls derived externally to the current clinical trial, to avoid the misconceptions that may arise from using other terms [19, 94].

3.2 | Emerging Definitions for Synthetic Data and RWD in EU, US, and UK Legal and Regulatory Frameworks

From the legal and regulatory framework, definitions have also emerged, both for synthetic data and RWD (the latter intended as a subset of observed/true data). The EU Data Governance Act (DGA) is the only legal text referring to synthetic data described as a "privacy-preserving method that could contribute to a more privacy-friendly processing of data" [95]. In line with our views, EU Policy texts are also referring to synthetic data (European Commission, 2024) along with data protection-specific sources, as "artificial data that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data" [96]. As part of the latest developments from the EU AI Act, synthetic data will be associated with so-called "general-purpose AI models" and there will be specific requirements in terms of risks and methodology (for generative AI systems and models) [97].

The EMA's, draft reflection paper on the use of AI in the medicinal product lifecycle mentions synthetic data as an instrument to "deploy differential privacy techniques" and for "increasing model performance" [98]. While the Medicines and Healthcare products Regulatory Agency (MHRA) does not have a formal position paper on the matter, a commentary authored by MHRA defined synthetic data as, "artificial data that mimic the properties of and relationships in real data" [99].

RWD are defined by EU legal texts as "health data generated outside of clinical studies" [100], a broad definition potentially encompassing both synthetic and non-synthetic data. The EMA and MHRA similarly define RWD as data relating to patient health status or delivery of health care collected outside of a clinical study/in routine clinical practice [101,102]. The FDA defined RWD as "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources" [103,104], while also providing in a recent commentary an useful distinction (in the context of true-observed data) between primary data collection and secondary data analysis, stressing the need for clarity in terminology of study designs [105].

With regards to synthetic data intended as artificially (hence as a product of a data-driven method [5]) generated, the regulatory landscape is adapting to this technology. As already stressed earlier, there is growing consideration for integrating novel methodologies like AI-generated synthetic data into the evidence generated to possibly support regulatory decision-making of medical products [48,98,106]. The MHRA and its Clinical Practice Research Datalink (CPRD) has been leading research efforts on synthetic data generation, including applications of high-fidelity synthetic data for purposes like validation of AI algorithms, data augmentation in the context of clinical trials for boosting sample sizes, and conditional boosting to address biases due to underrepresentation [107,108].

The EMA has shown a keen interest in the potential of AI, as shown in their reflection paper on AI in the medicinal product lifecycle, which acknowledges the significance of data augmentation techniques such as synthetic data in expanding training datasets for AI algorithms [98], while also stressing concepts such as generalizability and fairness of the models utilized/developed.

With regards to the US, the FDA has already started to recognize generative AI's potential, reporting the authorization of 1016 AI/ML-enabled medical devices as of December 2024 [109]. In the pharmaceutical realm, draft guidance on the use of AI in the drug development process have recently been published [106,110], however, no specific recommendations are outlined for the use of synthetic data or generative AI models in particular. Collectively, the regulatory agencies appear to be at various stages of recognizing and incorporating AI-generated (hence data-driven based [5]) synthetic data into their methodologies. All agencies appear to concur on the potential of synthetic data to enhance model performance and contribute to the medicinal product lifecycle, yet no drug or medical device has been registered using solely or predominantly synthetic data (as artificially generated data from a data-driven model) e.g., as a comparator arm [31]. It is reasonable to expect that quality aspects concerning synthetic data, possibly to be factored within statistical analyses, will need to be accounted for [111]. In fact, for the special case of predicting future outcomes given patient's baseline features (a use case often referred to as "digital twins"), a special case of ANCOVA has been referenced in regulatory discussions [112].

Besides new regulations and guidelines, it remains crucial for adopters of synthetic data to adhere to best practices and guidance documents related to data privacy, cybersecurity, and software validation in general. Moreover, given the current lack of an extensive guideline on the topic, the principles from the FDA's guidance's on the Use of Real-World Evidence may be applied to synthetic data, particularly in terms of ensuring data quality and reliability [103,104,106,109,113]. The underlying principles about software quality from the FDA's guidance on Computer Software Assurance for Production and Quality System Software can also be relevant to the algorithms and processes used to generate synthetic data [114], an approach that mitigates risks while also positioning organizations to adapt swiftly to new regulations as they are drafted and implemented.

In this context, following the recent issue of Good Machine Learning Practice for Medical Development[1] principles by FDA, MHRA and Health Canada (a source of information deemed general enough to also guide the application of AI/ML-methods in biopharmaceutical development [115] given the current absence of official comprehensive guidances) a first attempt to define some best practice for the development, evaluation and use of *in silico* methodologies—which, to varying extents, may involve the use or generation of synthetic data via data or process driven methods—is represented by the position report, "Toward Good Simulation Practice: Best Practices for the Use of Computational Modelling and Simulation in the Regulatory Process of Biomedical Products" [116]. The consensus process involved experts worldwide working in academia, healthcare, industry, and regulatory

bodies, including a team of 13 FDA computer modeling and simulation (CM&S) experts covering all three medical product centers: Center for Devices and Radiological Health (CDRH), Center for Drug Evaluation and Research (CDER), and Center for Biologics Evaluation and Research (CBER). Notably, the authors highlight in their report that current regulatory frameworks for assessing in silico methodologies do not align neatly with the traditional distinction between medicinal products and medical devices. According to the authors, these methodologies necessitate both elements of technical validation—typical of medical device regulatory pathways—and aspects of clinical validation, more commonly associated with medicinal product approvals. These initiatives are particularly interesting considering the key challenges associated with AI supported drug development endeavours highlighted by Nene et al. (2024) from the regulatory and sponsor perspectives [117]. On the regulatory side, difficulties such as inadequate description of data and insufficient evaluation or validation of models were emphasized. From the sponsor's viewpoint, unclear model requirements and insufficient guidance on relevant cases of interest were identified as major concerns. Addressing these challenges will likely be critical in improving the regulatory acceptance and practical implementation of innovative methodologies. Beyond the regulatory framework, it is also essential to consider the existing legal frameworks upon which the different authorities act. In the EU, for example, existing medical device and in vitro diagnostic regulations (MDR/IVDR) and pharmaceutical laws do not explicitly prohibit the use of synthetic data as a supporting element of clinical evidence [118,119]. For medical devices, the MDR per se does not prohibit using evidence generated through CM&S [120, 121], and therefore (artificially generated in terms of process-driven method [5]) synthetic data. For medicinal products, it is noteworthy that the EU pharmaceutical reform package even refers to "considering new approach methodologies in place of animal testing," including "in silico tools" [122], but in light of the related distinctions between synthetic data as generated by a more innovative data-driven (e.g., AI and related models) approach rather than an established process-driven one (e.g., CM&S as in QSP), more details and case examples will be required to determine their regulatory acceptance and differences.

In the presence of no explicit legal prohibitions, it is important that regulatory agencies and competent bodies take the initiative to align on these definitions and provide guidance on the use of synthetic data for generating evidence for medical products—so that healthcare stakeholders do not operate in a legal vacuum.

4 | Addressing Emerging Issues With Synthetic Data

As the application of synthetic data in healthcare and drug development continues to grow, several critical issues need to be addressed to ensure the data's reliability, provenance, and transparency [117].

4.1 | Provenance of Synthetic Data

Provenance, referring to the origin and history of data, provides a detailed record of its creation, transformation, and usage [123]. In the context of synthetic (as artificially generated) data, especially for the purpose of data augmentation, establishing robust provenance mechanisms is essential to maintain trust and credibility [102]. Unlike observed data, which usually has a clearer origin, synthetic data can be generated from algorithms and models that may combine multiple data sources [124] or mathematical models of an underlying biochemical process [5]. This complexity, as also proposed by The Data & Trust Alliance, necessitates detailed documentation of the models used, the observed ('real') data inputs (for data-driven generation processes), and the synthetic data generation methodology [125].

To tackle these challenges, developing comprehensive metadata standards is crucial. These standards should document the data generation process, including the algorithms used, parameters set, and input data characteristics [116]. Storing this metadata alongside the synthetic data provides context and supports reproducibility [126].

4.2 | Distinguishing Synthetic and Observed Data

As synthetic data becomes more integrated with observed data sources [49], distinguishing between the two is crucial to avoid misinterpretations and ensure appropriate usage in clinical research. The potential for data mixing, where synthetic and real data are combined, might pose significant challenges. Researchers may face difficulties identifying synthetic data elements, leading to potential biases or errors in analysis. Ensuring transparency in data usage and analysis is paramount, particularly when synthetic data augments observed datasets.

Adopting clear labeling practices where synthetic data are explicitly tagged and possible to be separated from observed data can mitigate these challenges. This may be achievable through data flags or markers embedded within the datasets. Furthermore, providing detailed documentation and visual aids, such as data lineage charts, may delineate the proportions and sources of synthetic and real data within mixed datasets. These practices could potentially enhance clarity and reduce the risk of misinterpretation.

4.2.1 | Developing Data Cards for Transparency

To address transparency issues, the concept of data cards has emerged as a potential solution [127]. Data cards are structured summaries that provide critical information about datasets, including their provenance, composition, and intended use. These cards may include detailed information about the original sources of the data, including any observed data sources (e.g., RWD, RCTs or mixtures thereof) used as input for generating synthetic data. Additionally, a description of the synthetic data generation process, including the algorithms and models employed, parameter settings, and any preprocessing steps, may be included [125].

Data cards may also highlight key characteristics of synthetic data, such as distributions, correlations, and outliers, through statistical summaries and visualizations [127].

Providing clear guidelines on the appropriate use of the data, potential limitations, and any known biases or uncertainties can strengthen the utility value of the data cards. Developing standardized templates for data cards that can be universally applied across different data and research projects will help ensure consistency. Leveraging automated tools to generate data cards as part of the synthetic data creation workflow can further enhance transparency and reduce manual effort.

4.3 | Data Synthesis and Replicability via Generative AI

The data-driven AI models that are utilized to produce synthetic data, while powerful, are imperfect and can fail to achieve their goals of producing high-quality synthetic data in a variety of ways [128], some of which we find important to highlight.

Firstly, synthetic data that are generated by necessity inherits the properties of the observed data distribution used to train an AI model, e.g., if only data from a specific demographic are shown to the model, the model will not produce data relevant to other demographics. This makes it important to carefully consider which data are used to train the model and whether it is consistent with the intended use of the synthetic data. Secondly, even if the original data contain all relevant demographics, they may not be reproduced by generative AI, as it may struggle to represent less frequent data [129]. Ensuring careful testing of synthetic data is vital to detect and address such issues [129]. Finally, a generative AI model may learn the data *too well*, i.e., it may memorize some real data points and output them under the guise of "synthetic" data [130]. This may endanger patient privacy and break relevant laws such as the General Data Protection Regulation (GDPR). Measuring privacy risks is an open research topic with no well-established procedures that nonetheless should not be ignored when applying generative AI in practice.

In terms of data replicability, El Emam et al. emphasize that the replicability of analyses performed on synthetic (as artificially generated) health data is a crucial factor in determining its validity for research or decision-making use [111]. Their study demonstrates that for synthetic data to yield reliable results, at least 10 datasets of the same size as the original should be generated and analyzed using multiple imputation combining rules. Moreover, the study highlights the superiority of sequential synthesis (a generative approach used to construct synthetic datasets by iteratively modeling each variable conditional on the previously synthesized ones, ensuring that dependencies among variables are preserved in the synthetic dataset) over GANs in replicating real-world analysis outcomes, ensuring high decision agreement, low bias, and appropriate confidence interval coverage.

Overall, a crucial aspect of evaluating the utility of synthetic data lies in its ability to yield conclusions that align with those derived from its original observed data source. If analyses conducted on synthetic and actual data lead to fundamentally

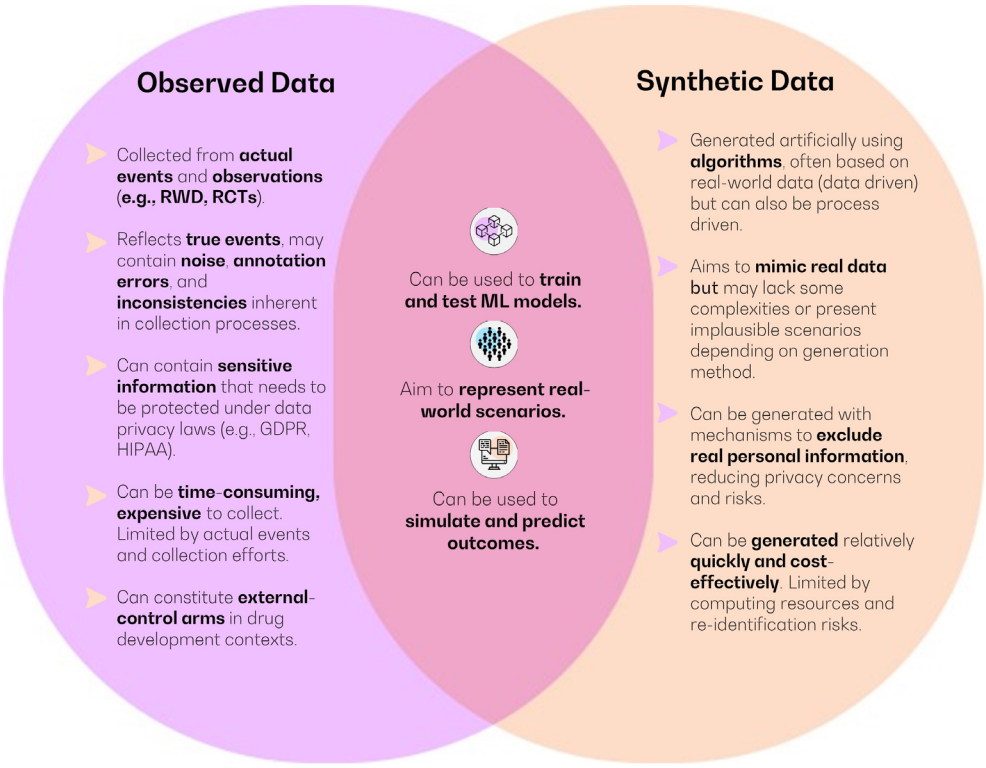

FIGURE 1 | Comparison of observed ("true") data and synthetic data via Venn diagram, illustrating the key characteristics, benefits, and limitations of each data type while highlighting their unique attributes and areas of overlap.

TABLE 1 | Terminology clarification and suggested definitions for conventions.

Term	Definition	Context/References
Observed data	Data that are obtained by direct measurement or collection from real-world events, are typically used as input to produce synthetic data. Includes RCT and RWD	Current paper and necessity to distinguish observed data from synthetic (as artificial) data [96]
Real-world data	Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. Can be further categorized into primary data (collected directly from study participants, either retrospectively or prospectively) and secondary data (obtained from existing healthcare data collection infrastructures, such as administrative claims databases, EHR databases, patient registries, or individual patient medical records)	FDA definition [134], and [64, 105, 135] for primary and secondary RWD distinction
Synthetic data (or fully synthetic data)	Data that have been created artificially (e.g., through statistical modeling, computer simulation) so that new values and/or data elements are generated	Current paper, FDA definition, and scientific discussions on the topic [3, 5, 49, 96]
Partially synthetic data	Data in which only selected variables are replaced with a synthetic (i.e., artificial) generated data values	Current paper, and as posed by authors from the US Department of Health and Human Services and Department of Health Administration and Policy [49]
Hybrid synthetic data	Data integrating observed ('real' or 'actual') and synthetic data	
External control arm	Control or comparator groups in clinical studies derived from external (concurrent) or historical sources of data. Depending on the nature of data used, it should be classified as a synthetic control arm if fully synthetic data are used, a partially synthetic control arm if only selected variables are replaced with synthetic data, or a hybrid synthetic control arm if records of observed and synthetic data are integrated	Current paper and as posed in regulatory contexts [12, 16, 17]; if synthetic data are included in an ECA, it should be explicitly reported whether the data generation method was process-driven or data-driven
Process-driven synthetic data	Data that have been created artificially using mechanistic or computational models that simulate biological/clinical processes	Used primarily in bioinformatics, computational biology, and clinical pharmacology (e.g., PBPK and popPKPD modeling) for simulating biological systems, widely recognized, regulated, and accepted by regulatory bodies [5]
Data-driven synthetic data	Data that have been created artificially after models have been trained on actual true/observed data	Typically derived from AI models trained on (true/observed) real-world or controlled datasets, increasingly employed for dataset augmentation, validation, or analytical purposes, yet currently less established within regulatory frameworks [5]

different conclusions, the synthetic dataset may lack validity for decision-making. Therefore, before releasing synthetic data for broader use, it is essential to assess its reliability through rigorous validation processes. This includes hypothesis testing and statistical analyses to ensure that key inferences—such as treatment effects or risk associations—remain consistent across synthetic and actual datasets. For such purpose, incorporation of model card, intended as a structured report detailing the technical characteristics of an AI model, benchmark evaluation results, and the context in which the model is designed to be used/methods employed to assess its performance [131] in such validation steps would enhance confidence in synthetic data as a viable tool for research and possibly regulatory decision-making.

5 | Conclusions

As the discourse on data provenance and classification unfolds within the clinical research and regulatory spheres, there is

an undeniable surge in the dialogue surrounding its two main identities found in literature: data derived from observed data sources (e.g., RWD or RCTs), and that generated by AI and other statistical (e.g., data-driven) means (artificial data), often and more frequently referred to as synthetic data [3,5,16,49]. As depicted in Figure 1, observed data and synthetic data possess distinct characteristics and potential applications, with certain overlapping benefits that enhance their utility in healthcare research and, potentially, drug development.

While there is an established regulatory basis for the former, evidenced by the integration of RWD or historical RCTs data into ECAs for upwards of 45 licensed drugs [17], the latter is still navigating its regulatory definition. As for data deriving from in silico and CM&S (the *process-driven* generated synthetic data [5]), the FDA, EMA, and MHRA routinely accept in silico evidence from physiologically based pharmacokinetic (PBPK), population pharmacokinetic (popPK) and/or pharmacodynamic (popPKPD), and exposure-response (ER) modeling in drug development [113,132,133]. Notwithstanding, the emergent generation of synthetic data through AI (the *data-driven* generated synthetic data [5]), as referenced in recent exploratory studies and regulatory endeavours, has yet to be defined in a harmonized fashion within regulatory frameworks, despite acknowledging the substantial potential of AI in this domain [48,98]. This dichotomy—between data rooted in actuality and data "born" from algorithms—highlights a need for clarity and consensus from stakeholders.

There is a compelling imperative for the health care and drug-development communities to define and distinguish these types of data and related derivations (see discussion on comparator arms in clinical trials and drug development in above sections) rigorously. The lack of clear distinction might dampen the current understanding and potential applications of these powerful tools, hence hampering the progress of possibly adopting AI-generated data within regulated drug development pathways. To aid in clarifying this terminology, we suggest (Table 1) the following conventions: We define observed data as ('real' or 'true') data that are obtained by direct measurement or collection from real-world events, (hence including RCT and RWD). We then define the term "real-world data" as defined by the FDA [134] as data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. Examples of RWD include data derived from EHRs, claims and billing data, data from product and disease registries, patient-generated data including in home-use settings, and data gathered from other sources that can inform on health status, such as mobile devices. RWD sources (e.g., registries, collections of EHRs, and administrative and healthcare claims databases) can be used as data collection and analysis infrastructure to support randomized controlled trials, including acting as an ECA source.

The term "synthetic" in the context of data is reserved for artificial data that are generated via algorithmic processes, as recently suggested in a regulatory context [3]. It is important to distinguish between synthetic data derived from process-driven and data-driven methods [5], as they rely on different underlying assumptions, technical frameworks, and regulatory considerations. *Process-driven* synthetic data, such as those generated through mechanistic modeling (e.g., QSP), have long been established and widely accepted in drug development, whereas *data-driven* synthetic data, often produced using AI and ML, remain relatively novel with no (to the best of our knowledge) regulatory precedent in the context of drug development/drug approval. The classification of synthetic data provided by Gonzales, Guruswamy and Smith [49] into fully, partially, or hybrid synthetic data is also useful to distinguish between subtypes of synthetic data. In the context of clinical trials using control or comparator arms drawn from RWD or historical controls from previous clinical trials, we suggest, in line with Seeger et al. [16], the use of the term "External Control Arm (ECA)" with specification of the source of the external controls (i.e., by means of observed data sources, generative AI techniques, or possible mixtures thereof).

In conclusion, it is evident that a collaborative dialogue among various communities, including academia, clinicians, industry, and regulatory advisors, can foster a shared understanding and guide the thoughtful exploration of synthetic data's potential in its finest declinations. It is through collective insights and expert discussions that the path forward can be envisioned, encouraging a harmonized perspective on synthetic data's role in advancing medical science and drug development.

Acknowledgments

We would like to extend our gratitude to Nataša Mandić (InSilicoTrials) for her contribution to the design of the image in this manuscript and Dr. Vinay Pai (FDA) for the insightful discussions on the topic. We also thank the three anonymous reviewers for their useful insights and comments, which greatly improved the quality of this work.

Conflicts of Interest

G.P., V.C., M.T., and D.R. are employees of InSilicoTrials Technologies S.p.A.; M.V. and A.C. are employees of InSilicoTrials Technologies B.V., two companies operating in modeling and simulation for drug development purposes. L.E. is the chief executive officer of InSilicoTrials Technologies S.p.A. A.V. is an equity-holding employee of Unlearn.AI Inc., a company that creates digital twin generators to forecast patient outcomes. C.F. is the chief executive officer of Unlearn.AI Inc. S.D. is the chief executive officer and chief technology officer of Train S.r.l., a company involved in the development of digital twin technology and synthetic data generation for precision medicine and drug development. The other authors declared no competing interests for this work.

Endnotes

[1] https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

References

1. European Commission, "European Health Data Space," (2023), https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en.

2. United States Census Bureau, "What Are Synthetic Data?," (2021), https://www.census.gov/about/what/synthetic-data.html.

3. FDA, "FDA Glossary on Digital Health and Artificial Intelligence," (2025), https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource#.

4. K. El Emam, "Status of Synthetic Data Generation for Structured Health Data," *JCO Clinical Cancer Informatics* 7 (2023): e2300071, https://doi.org/10.1200/cci.23.00071.

5. K. Selvarajoo and S. Maurer-Stroh, "Towards Multi-Omics Synthetic Data Integration," *Briefings in Bioinformatics* 25, no. 3 (2024): bbae213.

6. EMA, *Guideline on the Reporting of Physiologically Based Pharmacokinetic (PBPK) Modelling and Simulation* (European Union: Committee for Medicinal Products for Human Use (CHMP), 2018).

7. J. S. Owen and J. Fiedler-Kelly, *Introduction to Population Pharmacokinetic/Pharmacodynamic Analysis With Nonlinear Mixed Effects Models: Owen/Introduction to Population Pharmacokinetic/Pharmacodynamic Analysis With Nonlinear Mixed Effects Models* (John Wiley & Sons, Inc, 2014).

8. FDA, *Population Pharmacokinetics Guidance for Industry* (U.S. Department of Health and Human Services, 2022).

9. R. Upton and D. Mould, "Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development: Part 3-Introduction to Pharmacodynamic Modeling Methods," *CPT: Pharmacometrics & Systems Pharmacology* 3, no. 1 (2014): 88.

10. D. Mould and R. Upton, "Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development-Part 2: Introduction to Pharmacokinetic Modeling Methods," *CPT: Pharmacometrics & Systems Pharmacology* 2, no. 4 (2013): 38.

11. W. Wang, K. Hallow, and D. James, "A Tutorial on RxODE: Simulating Differential Equation Pharmacometric Models in R," *CPT: Pharmacometrics & Systems Pharmacology* 5, no. 1 (2016): 3–10.

12. FDA, *Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products* (U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Oncology Center of Excellence (OCE), 2023).

13. C. Serrano, S. Rothschild, G. Villacampa, et al., "Rethinking Placebos: Embracing Synthetic Control Arms in Clinical Trials for Rare Tumors," *Nature Medicine* 29, no. 11 (2023): 2689–2692.

14. J. M. Switchenko, A. L. Heeke, T. C. Pan, and W. L. Read, "The Use of a Predictive Statistical Model to Make a Virtual Control Arm for a Clinical Trial," *PLoS One* 14, no. 9 (2019): e0221336.

15. S. Kurki, V. Halla-Aho, M. Haussmann, H. Lähdesmäki, J. V. Leinonen, and M. Koskinen, "A Comparative Study of Clinical Trial and Real-World Data in Patients With Diabetic Kidney Disease," *Scientific Reports* 14, no. 1 (2024): 1731.

16. J. D. Seeger, K. J. Davis, M. R. Iannacone, et al., "Methods for External Control Groups for Single Arm Trials or Long-Term Uncontrolled Extensions to Randomized Clinical Trials," *Pharmacoepidemiology and Drug Safety* 29, no. 11 (2020): 1382–1392, https://doi.org/10.1002/pds.5141.

17. M. Jahanshahi, K. Gregg, G. Davis, et al., "The Use of External Controls in FDA Regulatory Decision Making," *Therapeutic Innovation & Regulatory Science* 55, no. 5 (2021): 1019–1035.

18. J. Lim, R. Walley, J. Yuan, et al., "Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities," *Therapeutic Innovation & Regulatory Science* 52, no. 5 (2018): 546–559.

19. X. Wang, F. Dormont, C. Lorenzato, A. Latouche, R. Hernandez, and R. Rouzier, "Current Perspectives for External Control Arms in Oncology Clinical Trials: Analysis of EMA Approvals 2016–2021," *Journal of Cancer Policy* 35 (2023): 100403.

20. S. D'Amico, D. Dall'Olio, C. Sala, et al., "Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology," *JCO Clinical Cancer Informatics* 7 (2023): e2300021.

21. J. M. Snowden, S. Rose, and K. M. Mortimer, "Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique," *American Journal of Epidemiology* 173, no. 7 (2011): 731–738.

22. A. Lehtinen and J. Raerinne, "Simulated Data in Empirical Science," *Foundations of Science* (2023): 2, https://doi.org/10.1007/s10699-023-09934-910.1007/s10699-023-09934-9.

23. W. Dong, D. Y. T. Fong, J. s. Yoon, et al., "Generative Adversarial Networks for Imputing Missing Data for Big Data Clinical Research," *BMC Medical Research Methodology* 21, no. 1 (2021): 78, https://doi.org/10.1186/s12874-021-01272-3.

24. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is all You Need," *Advances in Neural Information Processing Systems* 30 (2017): 3.

25. J. Hoffmann, S. Borgeaud, A. Mensch, et al., "Training Compute-Optimal Large Language Models," (2022), *ArXiv*, Prepr ArXiv220315556.

26. T. Brown, B. Mann, N. Ryder, et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020): 1877–1901.

27. X. Liu, F. Zhang, Z. Hou, et al., "Self-Supervised Learning: Generative or Contrastive," *IEEE Transactions on Knowledge and Data Engineering* 1–1 (2021): 1.

28. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139–144.

29. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, ed. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc, 2020), 6840–6851.

30. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," (2022), *arXiv*, [cited 2024 Mar 29], http://arxiv.org/abs/1312.6114.

31. C. Alloza, B. Knox, H. Raad, et al., "A Case for Synthetic Data in Regulatory Decision-Making in Europe," *Clinical Pharmacology and Therapeutics* 114, no. 4 (2023): 795–801.

32. G. Truda, "Generating Tabular Datasets Under Differential Privacy," (2023), *arXiv*, [cited 2024 Mar 26], http://arxiv.org/abs/2308.14784.

33. M. Goyal and Q. H. Mahmoud, "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI," *Electronics* 13, no. 17 (2024): 3509.

34. T. Kokosi, B. De Stavola, R. Mitra, et al., "An Overview on Synthetic Administrative Data for Research," *International Journal of Population Data Science* 7, no. 1 (2022): 3.

35. J. Noguer, I. Contreras, O. Mujahid, A. Beneyto, and J. Vehi, "Generation of Individualized Synthetic Data for Augmentation of the Type 1 Diabetes Data Sets Using Deep Learning Models," *Sensors* 22, no. 13 (2022): 4944.

36. F. Jacobs, S. D'Amico, C. Benvenuti, et al., "Opportunities and Challenges of Synthetic Data Generation in Oncology," *JCO Clinical Cancer Informatics* 7 (2023): e2300045.

37. H. Y. J. Kang, E. Batbaatar, D. W. Choi, K. S. Choi, M. Ko, and K. S. Ryu, "Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy," *JMIR Medical Informatics* 11 (2023): e47859.

38. R. F. Baumeister and M. R. Leary, "Writing Narrative Literature Reviews," *Review of General Psychology* 1, no. 3 (1997): 311–320.

39. A. Sutton, M. Clowes, L. Preston, and A. Booth, "Meeting the Review Family: Exploring Review Types and Associated Information Retrieval Requirements," *Health Information and Libraries Journal* 36, no. 3 (2019): 202–222.

40. Y. Cheng, R. Straube, A. E. Alnaif, L. Huang, T. A. Leil, and B. J. Schmidt, "Virtual Populations for Quantitative Systems Pharmacology Models," *Methods in Molecular Biology* 2486 (2022): 129–179.

41. F. Umer and N. Adnan, "Generative Artificial Intelligence: Synthetic Datasets in Dentistry," *BDJ Open* 10, no. 1 (2024): 13.

42. A. Arora and A. Arora, "Machine Learning Models Trained on Synthetic Datasets of Multiple Sample Sizes for the Use of Predicting Blood Pressure From Clinical Data in a National Dataset," *PLoS One* 18, no. 3 (2023): e0283094.

43. B. Baudry, K. Etemadi, S. Fang, et al., "Generative AI to Generate Test Data Generators," (2024), [cited 2024 Mar 26], https://arxiv.org/abs/2401.17626.

44. M. Herper, J. Apteakar, and A. Shafquat, "Fake Data, Real Insights: How Virtual Patients Could Transform Clinical Trials," (2024), https://www.statnews.com/sponsor/2024/01/24/fake-data-real-insights-how-virtual-patients-could-transform-clinical-trials/.

45. C. K. Fisher, A. M. Smith, J. R. Walsh, et al., "Machine Learning for Comprehensive Forecasting of Alzheimer's Disease Progression," *Scientific Reports* 9, no. 1 (2019): 13622.

46. Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, and K. El Emam, "Can Synthetic Data Be a Proxy for Real Clinical Trial Data? A Validation Study," *BMJ Open* 11, no. 4 (2021): e043497.

47. S. El Kababji, N. Mitsakakis, X. Fang, et al., "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets," *JCO Clinical Cancer Informatics* 7 (2023): e2300116.

48. P. Myles, J. Ordish, and A. Tucker, "The Potential Synergies Between Synthetic Data and In Silico Trials in Relation to Generating Representative Virtual Population Cohorts," *Progress in Biomedical Engineering* 5, no. 1 (2023): 013001.

49. A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic Data in Health Care: A Narrative Review," *PLOS Digital Health* 2, no. 1 (2023): e0000082.

50. D. J. Boyne, D. E. Dawe, H. Shakir, et al., "Comparative Effectiveness of Lurbinectedin for the Treatment of Relapsed Small Cell Lung Cancer in the Post-Platinum Setting: A Real-World Canadian Synthetic Control Arm Analysis," *Targeted Oncology* 18, no. 5 (2023): 697–705.

51. H. Van Le, K. Van Naarden Braun, G. S. Nowakowski, et al., "Use of a Real-World Synthetic Control Arm for Direct Comparison of Lisocabtagene Maraleucel and Conventional Therapy in Relapsed/Refractory Large B-Cell Lymphoma," *Leukemia & Lymphoma* 64, no. 3 (2023): 573–585.

52. C. A. W. Bruhn, S. Hetterich, C. Schuck-Paim, et al., "Estimating the Population-Level Impact of Vaccines Using Synthetic Controls," *Proceedings of the National Academy of Sciences* 114, no. 7 (2017): 1524–1529.

53. M. Burcu, N. A. Dreyer, J. M. Franklin, et al., "Real-World Evidence to Support Regulatory Decision-Making for Medicines: Considerations for External Control Arms," *Pharmacoepidemiology and Drug Safety* 29, no. 10 (2020): 1228–1235.

54. R. Banerjee, S. Midha, A. H. Kelkar, A. Goodman, V. Prasad, and G. R. Mohyuddin, "Synthetic Control Arms in Studies of Multiple Myeloma and Diffuse Large B-Cell Lymphoma," *British Journal of Haematology* 196, no. 5 (2022): 1274–1277.

55. S. O'Haire, K. Degeling, F. Franchini, et al., "Comparing Survival Outcomes for Advanced Cancer Patients Who Received Complex Genomic Profiling Using a Synthetic Control Arm," *Targeted Oncology* 17, no. 5 (2022): 539–548.

56. S. Popat, S. V. Liu, N. Scheuer, et al., "Addressing Challenges With Real-World Synthetic Control Arms to Demonstrate the Comparative Effectiveness of Pralsetinib in Non-Small Cell Lung Cancer," *Nature Communications* 13, no. 1 (2022): 3500.

57. J. Zhu and R. S. Tang, "A Proper Statistical Inference Framework to Compare Clinical Trial and Real-World Progression-Free Survival Data," *Statistics in Medicine* 41, no. 29 (2022): 5738–5752.

58. N. Neehal, V. Anand, and K. P. Bennett, "Framework for Research in Equitable Synthetic Control Arms," *AMIA Annual Symposium Proceedings/AMIA Symposium* 2023 (2023): 530–539.

59. T. Yoshino, Q. Shi, T. Misumi, et al., "A Synthetic Control Arm for Refractory Metastatic Colorectal Cancer: The no Placebo Initiative," *Nature Medicine* 29, no. 10 (2023): 2389–2390.

60. K. Thorlund, S. Duffield, S. Popat, et al., "Quantitative Bias Analysis for External Control Arms Using Real-World Data in Clinical Trials: A Primer for Clinical Researchers," *Journal of Comparative Effectiveness Research* 13, no. 3 (2024): e230147.

61. Y. Uemura, R. Ozaki, T. Shinozaki, et al., "Comparative Effectiveness of Tocilizumab vs Standard Care in Patients With Severe COVID-19-Related Pneumonia: A Retrospective Cohort Study Utilizing Registry Data as a Synthetic Control," *BMC Infectious Diseases* 23, no. 1 (2023): 849.

62. M. E. Menefee, Y. Gong, P. S. Mishra-Kalyani, et al., "Project Switch: Docetaxel as a Potential Synthetic Control in Metastatic Non-Small Cell Lung Cancer (mNSCLC) Trials," *Journal of Clinical Oncology* 37, no. 15_suppl (2019): 9105.

63. B. Walker, H. E. Ray, P. Shao, C. D'Ambrosio, C. White, and M. S. Walker, "Comparing Prospectively Assigned Trial and Real-World Lung Cancer Patients," *Journal of Comparative Effectiveness Research* 13, no. 7 (2024): e230176.

64. G. Prada-Ramallal, F. Roque, M. T. Herdeiro, B. Takkouche, and A. Figueiras, "Primary Versus Secondary Source of Data in Observational Studies and Heterogeneity in Meta-Analyses of Drug Effects: A Survey of Major Medical Journals," *BMC medical research methodology* 18, no. 1 (2018): 97, https://doi.org/10.1186/s12874-018-0561-3.

65. R. Davi, M. Chandler, B. Elashoff, et al., "Non-Small Cell Lung Cancer (NSCLC) Case Study Examining Whether Results in a Randomized Control Arm Are Replicated by a Synthetic Control Arm (SCA)," *Journal of Clinical Oncology* 37, no. 15_suppl (2019): 9108.

66. R. Davi, X. Yin, and M. Stewart, "Exploring the Validity of a Synthetic Control Arm (SCA) for Augmentation or Replacement of a Randomized Control in Difficult-To-Study Indications: A Case Study in Relapsed or Refractory Multiple Myeloma (R/R MM)," *Journal of Clinical Oncology* 38, no. 15_suppl (2020): e20521.

67. H. Li, R. Tiwari, and Q. H. Li, "Conditional Borrowing External Data to Establish a Hybrid Control Arm in Randomized Clinical Trials," *Journal of Biopharmaceutical Statistics* 32, no. 6 (2022): 954–968.

68. W. K. Tan, B. D. Segal, M. D. Curtis, et al., "Augmenting Control Arms With Real-World Data for Cancer Trials: Hybrid Control Arm Methods and Considerations," *Contemporary Clinical Trials Communications* 30 (2022): 101000.

69. S. Sengupta, I. Ntambwe, K. Tan, et al., "Emulating Randomized Controlled Trials With Hybrid Control Arms in Oncology: A Case Study," *Clinical Pharmacology and Therapeutics* 113, no. 4 (2023): 867–877.

70. K. H. Zou, C. Vigna, A. Talwai, et al., "The Next Horizon of Drug Development: External Control Arms and Innovative Tools to Enrich Clinical Trial Data," *Therapeutic Innovation & Regulatory Science* 58, no. 3 (2024): 443–455.

71. G. Carrigan, S. Whipple, W. B. Capra, et al., "Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-Of-Concept in Randomized Controlled Trials," *Clinical Pharmacology and Therapeutics* 107, no. 2 (2020): 369–377.

72. D. Backenroth, "How to Choose a Time Zero for Patients in External Control Arms," *Pharmaceutical Statistics* 20, no. 4 (2021): 783–792.

73. J. R. Curtis, V. Strand, S. Golombek, et al., "Patient Outcomes Improve When a Molecular Signature Test Guides Treatment Decision-Making in Rheumatoid Arthritis," *Expert Review of Molecular Diagnostics* 22, no. 10 (2022): 1–10.

74. S. Børø, S. Thoresen, S. Boge Brant, and Å. Helland, "Initial Investigation of Using Norwegian Health Data for the Purpose of External Comparator Arms – An Example for Non-Small Cell Lung Cancer," *Acta Oncologica* 62, no. 12 (2023): 1642–1648.

75. M. Carton, J. P. Del Castillo, J. B. Colin, et al., "Larotrectinib Versus Historical Standard of Care in Patients With Infantile Fibrosarcoma: Protocol of EPI-VITRAKVI," *Future Oncology (London, England)* 19, no. 24 (2023): 1645–1653.

76. L. E. Dang, E. Fong, J. M. Tarp, et al., "Case Study of Semaglutide and Cardiovascular Outcomes: An Application of the C *ausal Roadmap* to a Hybrid Design for Augmenting an RCT Control Arm With Real-World Data," *Journal of Clinical and Translational Science* 7, no. 1 (2023): e231.

77. S. Goulden, Q. Shen, R. L. Coleman, et al., "Outcomes for Dostarlimab and Real-World Treatments in Post-Platinum Patients With Advanced/Recurrent Endometrial Cancer: The GARNET Trial Versus a US Electronic Health Record-Based External Control Arm," *Journal of Health Economics and Outcomes Research* 10, no. 2 (2023): 53–61.

78. C. Abé, J. Keto, M. Lilja, et al., "Cytarabine Dose Intensification Improves Survival in Older Patients With Secondary/High-Risk Acute Myeloid Leukemia in Matched Real-World Versus Clinical Trial Data," *Leukemia & Lymphoma* 65, no. 10 (2024): 1493–1501.

79. E. Farah, M. Kenney, M. T. Warkentin, W. Y. Cheung, and D. R. Brenner, "Examining External Control Arms in Oncology: A Scoping Review of Applications to Date," *Cancer Medicine* 13, no. 13 (2024): e7447.

80. C. Gray, E. Ralphs, M. P. Fox, et al., "Use of Quantitative Bias Analysis to Evaluate Single-Arm Trials With Real-World Data External Controls," *Pharmacoepidemiology and Drug Safety* 33, no. 5 (2024): e5796.

81. J. M. Strayhorn, "Virtual Controls as an Alternative to Randomized Controlled Trials for Assessing Efficacy of Interventions," *BMC Medical Research Methodology* 21, no. 1 (2021): 3.

82. Stanford University, "Virtual Control Arms for Clinical Trials Using Deep Learning," (2023), https://techfinder.stanford.edu/technology/virtual-control-arms-clinical-trials-using-deep-learning.

83. K. Nicholson, J. Chan, E. A. Macklin, et al., "Pilot Trial of Inosine to Elevate Urate Levels in Amyotrophic Lateral Sclerosis," *Annals of Clinical Translational Neurology* 5, no. 12 (2018): 1522–1533.

84. Z. Chen, H. Zhang, Y. Guo, et al., "Exploring the Feasibility of Using Real-World Data From a Large Clinical Data Research Network to Simulate Clinical Trials of Alzheimer's Disease," *NPJ Digital Medicine* 4, no. 1 (2021): 84.

85. H. Folse, C. Sternhufvud, C. Andy Schuetz, B. Rengarajan, and S. Gandhi, "Impact of Switching Treatment From Rosuvastatin to Atorvastatin on Rates of Cardiovascular Events," *Clinical Therapeutics* 36, no. 1 (2014): 58–69.

86. R. Visentin, C. Dalla Man, B. Kovatchev, and C. Cobelli, "The University of Virginia/Padova Type 1 Diabetes Simulator Matches the Glucose Traces of a Clinical Trial," *Diabetes Technology & Therapeutics* 16, no. 7 (2014): 428–434.

87. R. Dutta, A. Mohan, J. Buros-Novik, G. Goldmacher, O. O. Akala, and B. Topp, "A Bootstrapping Method to Optimize Go/No-Go Decisions From Single-Arm, Signal-Finding Studies in Oncology," *CPT: Pharmacometrics & Systems Pharmacology* 13, no. 8 (2024): 1317–1326.

88. N. Loiseau, P. Trichelair, M. He, et al., "External Control Arm Analysis: An Evaluation of Propensity Score Approaches, G-Computation, and Doubly Debiased Machine Learning," *BMC Medical Research Methodology* 22, no. 1 (2022): 335.

89. P. M. McMahon, C. Y. Kong, B. E. Johnson, et al., "Estimating Long-Term Effectiveness of Lung Cancer Screening in the Mayo CT Screening Study," *Radiology* 248, no. 1 (2008): 278–287, https://doi.org/10.1148/radiol.2481071446.

90. S. Suissa, "Single-Arm Trials With Historical Controls: Study Designs to Avoid Time-Related Biases," *Epidemiology* 32, no. 1 (2021): 94–100.

91. J. Davies, M. Martinec, P. Delmar, et al., "Comparative Effectiveness From a Single-Arm Trial and Real-World Data: Alectinib Versus Ceritinib," *Journal of Comparative Effectiveness Research* 7, no. 9 (2018): 855–865.

92. P. S. Mishra-Kalyani, L. Amiri Kordestani, D. R. Rivera, et al., "External Control Arms in Oncology: Current Use and Future Directions," *Annals of Oncology* 33, no. 4 (2022): 376–383, https://doi.org/10.1016/j.annonc.2021.12.015.

93. A. Zayadi, R. Edge, C. E. Parker, et al., "Use of External Control Arms in Immune-Mediated Inflammatory Diseases: A Systematic Review," *BMJ Open* 13, no. 12 (2023): e076677.

94. International Conference on Harmonization (ICH), "E10: Choice of Control Group and Related Issues in Clinical Trials," (2000), https://database.ich.org/sites/default/files/E10_Guideline.pdf.

95. EU Parliament and Council, "Regulation (EU) 2022/868 on European Data Governance and Amending Regulation (EU) 2018/1724," *Official Journal of the European Union* 7 (2022): L152/5.

96. European Data Protection Supervisor, "Synthetic Data," (2022), https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en.

97. European Parliament, *EU AI Act: First Regulation on Artificial Intelligence* (European Parliament, 2023), https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

98. EMA, "Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle," in *European Union: Committee for Medicinal Products for Veterinary Use (CVMP)* (European Medicines Agency, 2023).

99. P. Myles, J. Ordish, and R. Branson, "Synthetic Data and the Innovation, Assessment, and Regulation of AI Medical Devices," *RF Quarterly* 2, no. 4 (2022): 20–26.

100. EU Parliament and Council, "Regulation (EU) 2022/123 on a Reinforced Role for the European Medicines Agency in Crisis Preparedness and Management for Medicinal Products and Medical Devices," *Official Journal of the European Union* 45 (2022): L20/09.

101. MHRA, "MHRA Guidance on the Use of Real-World Data in Clinical Studies to Support Regulatory Decisions," (2021), https://www.gov.uk/government/publications/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions.

102. EMA, *Data Quality Framework for EU Medicines Regulation: Application to Real-World Data* (Committee for Medicinal Products for Human Use (CHMP), 2024), https://www.ema.europa.eu/.

103. FDA, "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff," (2017).

104. FDA, "Framework for FDA's Real-World Evidence Program," (2018), https://www.fda.gov/media/120060/download.

105. J. Concato, P. Stein, G. J. Dal Pan, R. Ball, and J. Corrigan-Curay, "Randomized, observational, interventional, and real-world-What's in a name?," *Pharmacoepidemiology and drug safety* 29, no. 11 (2020): 1514–1517, https://doi.org/10.1002/pds.5123.

106. FDA, *Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products* (U.S. Department of Health and Human Services, Food and Drug Administration, 2023), 30313–30314.

107. CPRD, "Synthetic Data," (2021), https://www.cprd.com/content/synthetic-data.

108. Forrester, "AI 2.0: Upgrade Your Enterprise With Five Next-Generation AI Advances," (2021), https://www.forrester.com/report/AI-20-Upgrade-Your-Enterprise-With-Five-NextGeneration-AI-Advances/RES163520.

109. FDA, *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices* (U.S. Food and Drug Administration, 2023), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

110. FDA, *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry and Other Interested Parties* (Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH), Center for Veterinary Medicine (CVM), Oncology Center of Excellence (OCE), Office of Combination Products (OCP), Office of Inspections and Investigations (OII), 2025), https://www.regulations.gov.

111. K. El Emam, L. Mosquera, X. Fang, and A. El-Hussuna, "An Evaluation of the Replicability of Analyses Using Synthetic Health Data," *Scientific Reports* 14, no. 1 (2024): 6978.

112. Unlearn.AI, "Request for CHMP Qualification for Prognostic Covariate Adjustment (PROCOVA™) as an Efficient Statistical Methodology, Intended to Improve the Efficiency of Phase 2 and 3 Clinical Trials by Using Trial Subjects' Predicted Control Outcomes (Prognostic Scores) in Linear Covariate Adjustment," (2021), https://www.ema.europa.eu/en/documents/other/briefing-book_en.pdf.

113. FDA, *Successes and Opportunities in Modeling & Simulation for FDA* (U.S. Food and Drug Administration, 2022).

114. FDA, "Computer Software Assurance for Production and Quality System Software," (2022).

115. K. Köchert, T. Friede, M. Kunz, H. Pang, Y. Zhou, and E. Rantou, "On the Application of Artificial Intelligence/Machine Learning (AI/ML) in Late-Stage Clinical Development," *Therapeutic innovation & regulatory science* 58, no. 6 (2024): 1080–1093, https://doi.org/10.1007/s43441-024-00689-4.

116. M. Viceconti and L. Emili, eds., *Toward Good Simulation Practice: Best Practices for the Use of Computational Modelling and Simulation in the Regulatory Process of Biomedical Products* (Springer, 2024), https://link.springer.com/book/10.1007/978-3-031-48284-7.

117. L. Nene, B. T. Flepisi, S. J. Brand, C. Basson, and M. Balmith, "Evolution of Drug Development and Regulatory Affairs: The Demonstrated Power of Artificial Intelligence," *Clinical therapeutics* 46, no. 8 (2024): e6–e14, https://doi.org/10.1016/j.clinthera.2024.05.012.

118. EU Parliament and Council, "Regulation (EU) 2017/745 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EEC," *Official Journal of the European Union* L 117 (2017): 1–175.

119. EU Parliament and Council, "Regulation (EU) 2017/746 on In Vitro Diagnostic Medical Devices and Repealing Directive 98/79/EC and Commission Decision 2010/227/EU," *Official Journal of the European Union* L 117 (2017): 176–332.

120. E. Biasin, B. Yasar, and E. Kamenjašević, "New Cybersecurity Requirements for Medical Devices in the EU: The Forthcoming European Health Data Space, Data Act, and Artificial Intelligence Act," *Law, Technology and Humans* 5, no. 2 (2023): 43–58.

121. E. Biasin, M. Viceconti, I. Carbone, et al., "In Silico World D9.4 Validation and Recommendations for Lawmakers and Policymakers (Version 1)," *Zenodo* (2024), https://doi.org/10.5281/zenodo.14282240.

122. EU Commission, "Boosting Startups and Innovation in Trustworthy Artificial Intelligence," Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, (2024).

123. M. Johns, T. Meurers, F. N. Wirth, et al., "Data Provenance in Biomedical Research: Scoping Review," *Journal of Medical Internet Research* 25 (2023): e42289.

124. C. Little, M. Elliot, and R. Allmendinger, "Federated Learning for Generating Synthetic Data: A Scoping Review," *International Journal of Population Data Science* 8, no. 1 (2023): 2158.

125. Data & Trust Alliance, *Data Provenance Standards* (Data & Trust Alliance, 2024), https://dataandtrustalliance.org/our-initiatives/data-provenance-standards.

126. J. Leipzig, D. Nüst, C. T. Hoyt, K. Ram, and J. Greenberg, "The Role of Metadata in Reproducible Computational Research," *Patterns* 2, no. 9 (2021): 100322.

127. M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI," in *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022), 1776–1826.

128. H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, and A. Bano, "Synthetic Data Generation: State of the Art in Health Care Domain," *Computer Science Review* 48 (2023): 100546.

129. K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett, "The Problem of Fairness in Synthetic Healthcare Data," *Entropy* 23, no. 9 (2021): 1165.

130. C. Song, T. Ristenpart, and V. Shmatikov, "Machine Learning Models That Remember Too Much," (2017), *arXiv*, [cited 2025 Jan 31], https://arxiv.org/abs/1709.07886.

131. M. Mitchell, S. Wu, A. Zaldivar, et al., "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2019), 220–229, https://doi.org/10.1145/3287560.3287596.

132. T. Shepard, "Role of Modelling and Simulation in Regulatory Decision Making in Europe," London (2011).

133. E. Redrup, C. Mitchell, P. Myles, R. Branson, and A. F. Frangi, "Cross-Regulator Workshop: Journeys, Experiences and Best Practices on Computer Modelled and Simulated Regulatory Evidence—Workshop Report," (2023), https://doi.org/10.5281/zenodo.10121103.

134. FDA, "Real World Evidence," (2022), https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

135. S. Schneeweiss, and E. Patorno, "Conducting Real-world Evidence Studies on the Clinical Outcomes of Diabetes Treatments." *Endocrine reviews*, 42, no.5 (2021): 658–690. https://doi.org/10.1210/endrev/bnab007.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.