Virtual Control Arms in Clinical Trials: Bias & Validation

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · virtual control arms · external control arm · synthetic control arm · single-arm trial · clinical trials · real world data · FDA guidance · regulatory science · sensitivity analysis · exchangeability

Executive Summary

An external control arm (ECA) is a comparator group assembled from data outside a clinical trial, historical registries, electronic health record (EHR) networks, prior trials, or, in the newest variant, artificial intelligence (AI) generated synthetic patients, used when a single-arm trial (SAT) has no randomized comparator of its own. The U.S. Food and Drug Administration (FDA) defines an externally controlled trial as one comparing outcomes for treated participants against a group external to the trial, drawn from an earlier time (historical) or a concurrent different setting. FDA issued a dedicated draft guidance in February 2023, warning that the design is inappropriate when the expected treatment effect is modest because residual bias can distort the result (www.fda.gov). The European Medicines Agency (EMA) adopted a related single-arm-trial reflection paper in September 2024 (www.ema.europa.eu), and, as of May 2026, still describes dedicated European Union guidance on external controls as lacking in a new concept paper (www.ema.europa.eu), a real transatlantic asymmetry.

FDA's NDA/BLA study-design table reports two externally controlled trials in fiscal year (FY) 2023, zero in FY2024, and four in FY2025; each NDA/BLA is reported in the fiscal year in which FDA took regulatory action (www.fda.gov). Bibliometric reviews found 17% of EMA-approved cancer drugs from 2016 to 2021 used external controls (www.ebi.ac.uk) and that 47% of FDA oncology accelerated approvals between 1992 and 2020 relied on a single-arm trial (www.ebi.ac.uk). The defensibility of any external control arm rests on exchangeability, the degree to which trial and external patients are truly comparable, which adjustment methods such as propensity score matching and matching-adjusted indirect comparison approximate but cannot guarantee.

FDA's February 2023 draft guidance says sensitivity analyses should be used to evaluate the potential impact of plausible violations in missing-data assumptions on key analyses (FDA draft guidance). The strongest evidence of fitness for purpose is prospective-style validation: an EHR-derived external control methodology reproduced a known randomized hazard ratio in ten of eleven tested analyses in non-small-cell lung cancer (www.rtihs.org).

Documented case studies span registry-based historical controls for Duchenne muscular dystrophy (eteplirsen) (www.accessdata.fda.gov) to a 694-patient weighted historical cohort for blinatumomab in acute lymphoblastic leukemia (pmc.ncbi.nlm.nih.gov). Sponsors evaluating an external control arm should treat it as a validation exercise first and a statistical exercise second, prespecifying sensitivity analyses and seeking prospective validation wherever a reference trial exists. Readers seeking the AI-generated digital-twin sub-case in depth should consult the companion IntuitionLabs report on digital twins as virtual control arms (intuitionlabs.ai), which this report deliberately does not duplicate.

Introduction and Background

A single-arm trial (SAT), in which every enrolled patient receives the investigational treatment with no randomized comparator, cannot on its own demonstrate that an observed benefit is due to the treatment rather than the natural course of disease. Sponsors and regulators have long supplemented such trials with an external control arm (ECA): a group of patients who did not receive the treatment under evaluation, assembled from sources outside the trial itself, such as historical registries, prior clinical trials, or electronic health record (EHR) data. The U.S. Food and Drug Administration (FDA) defines an externally controlled trial as one in which outcomes for participants receiving the test treatment are compared against outcomes in a group external to the trial, drawn either from an earlier time period (a historical control) or from the same time period in a different setting (a concurrent external control) (www.fda.gov).

The terminology in this space is inconsistent across sources. "Historical control," "external control arm," "synthetic control arm," and "virtual control arm" are often used interchangeably in practice, though FDA guidance treats the historical control as one recognized subset of the broader external-control category. This report uses "external control arm" as the general term for any non-randomized comparator built from data outside the trial, and reserves "synthetic" or "virtual" control arm for the specific, newer sub-case in which artificial intelligence (AI) models generate simulated patient-level outcomes rather than drawing on directly observed comparator patients. A companion IntuitionLabs report examines that AI-generated sub-case in depth, covering digital twins as virtual control arms in detail (intuitionlabs.ai); this report instead concentrates on the broader methodological core shared by all external control arm designs, whether built from registries, EHR networks, prior trials, or generative models: exchangeability, bias, leakage-adjacent data quality risks, sensitivity analysis, and prospective validation.

That methodological core has become urgent because regulatory interest has accelerated markedly. FDA issued a dedicated draft guidance on externally controlled trials in February 2023, the European Medicines Agency (EMA) adopted a reflection paper on single-arm trials in September 2024 (www.ema.europa.eu), and, as of May 2026, EMA has published a concept paper explicitly acknowledging that specific European Union (EU) guidance on external controls is still lacking, ahead of a forthcoming reflection paper (www.ema.europa.eu). This report works through that regulatory landscape, the data sources and vendors that populate external control arms, the statistical machinery used to build and defend them, the validation and sensitivity-analysis frameworks regulators expect, and documented case studies, before closing with implications for sponsors weighing this design against a traditional randomized controlled trial (RCT). All facts below are dated at the source and were verified in September 2026 against the cited page.

Key Changes

External Control Arms

FDA's guidance cross-references the International Council for Harmonisation (ICH) E10 guideline, which addresses the choice of control group in clinical trials and discusses five principal types of controls (www.fda.gov). The February 2023 draft guidance goes further by addressing modern real-world data (RWD) sources directly, stating that the ability to control for confounding is fundamentally limited because a disease's natural history is often poorly understood and fit-for-use data on suspected confounders may be missing for some patients (www.fda.gov). It further warns that when the anticipated treatment effect is modest, an externally controlled design may not be appropriate because of concerns that residual bias will distort the comparison (www.fda.gov), and it states that sponsors must include relevant patient-level data for both treatment and external control arms in marketing applications, as required under FDA regulations.

FDA's real-world evidence program has been building toward this position for years. FDA first approved Blincyto under accelerated approval in December 2014 for Philadelphia chromosome-negative relapsed or refractory B-cell precursor acute lymphoblastic leukemia; regular approval for that indication was granted on July 11, 2017, when the indication was expanded to include Philadelphia chromosome-positive disease (FDA). FDA's December 2018 framework described the original approval's single-arm trial and historical comparison with 694 comparable patients drawn from more than 2,000 records, while noting difficulties in selecting comparable populations (www.fda.gov).

EMA's parallel work starts from single-arm trials rather than external controls specifically. Its September 2024 reflection paper states that formal proof that a single-arm trial's treatment-effect estimate is unbiased is impossible, even after bias-mitigation strategies are applied (www.ema.europa.eu), but does not give full treatment to designs that prospectively add a non-randomized external control arm. A May 2026 EMA concept paper is meant to close that gap, committing the agency to a forthcoming reflection paper addressing external-control terminology (including "external control," "synthetic data," and "digital twins"), data quality dimensions, and the ICH E9(R1) estimand framework together with target trial emulation. As of September 2026, a sponsor seeking EU acceptance of an external control arm faces FDA's February 2023 draft guidance but only an EMA concept paper promising future guidance, a genuine transatlantic asymmetry.

Real-World Data Foundations: Registries, EHR Networks, and Specialist Vendors

Every external control arm is only as good as the data feeding it, and a small set of data sources and vendors has emerged to supply that data at scale. Flatiron Health, an oncology-focused EHR data company, states on its own site that its real-world evidence has informed more than 45 global regulatory decisions (flatiron.com), a first-party vendor claim rather than an independently audited count. Komodo Health describes its data foundation as linking more than one trillion records spanning event data from more than 330 million de-identified patients, refreshed from more than 60 sources (www.komodohealth.com), and TriNetX positions its offering as a federated network connecting researchers directly to real-world health data drawn from participating providers' own EHR systems (trinetx.com). All three figures are vendor-stated scale claims, current as of September 2026, and should be read as such rather than as independently verified statistics.

On the services side, ConcertAI markets purpose-built external control arms and states, as a vendor claim, that they can reduce patient accrual needs in single-arm oncology trials by roughly half (www.concertai.com), while Verana Health markets curated historical cohorts (veranahealth.com). IQVIA cites its own HTA Accelerator analysis finding that submissions relying on a pivotal single-arm trial are more likely to receive a positive HTA decision when paired with an RWD-based external comparator (www.iqvia.com), a vendor-produced finding best read alongside independent literature.

Independent methodology work identifies challenges common to all of these sources: selecting suitable RWD to contextualize a single-arm trial requires resolving mismatches in how patients are identified, outcomes defined, and time anchored between the trial population and the external dataset. A systematic review of twenty qualifying FDA rare-disease NDA and BLA approvals found that external RWD controls drawn from other studies made up only a small share of those applications (two of twenty, or 10%), with most relying instead on natural-history or registry-based historical controls (www.ncbi.nlm.nih.gov).

The Methodological Core: Exchangeability, Confounding, and Statistical Adjustment

The central concept governing whether an external control arm is defensible is exchangeability: trial patients and external comparator patients must be similar enough, across eligibility, baseline characteristics, treatment definitions, outcome measurement, time period, and setting, that any outcome difference can be attributed to treatment rather than to who the two groups were. Peer-reviewed methodology literature is explicit that external validity is compromised whenever patient characteristics vary substantially between the two populations (pmc.ncbi.nlm.nih.gov), and that statistical methods including propensity score matching (PSM) are increasingly applied specifically to make trial and external-comparator subjects as similar as possible on observed characteristics (pmc.ncbi.nlm.nih.gov).

PSM and its relative, inverse probability of treatment weighting (IPTW), are not the only adjustment tools in use. A matching-adjusted indirect comparison (MAIC) reweights individual patient data from one arm so its aggregate characteristics match a comparator population's published summary statistics; Bayesian dynamic borrowing offers an alternative family of methods, including power priors, commensurate priors, meta-analytic predictive priors, and robust mixture priors, that let a model down-weight external data according to how much it appears to conflict with the trial's own concurrent evidence (pmc.ncbi.nlm.nih.gov). These methods can fail even when applied carefully.

Even a well-matched external control arm carries structural bias risks that adjustment alone cannot fix. When trial and external-comparator patients differ in the timing of treatment decisions, the comparison can introduce immortal time bias or time-lag bias, since patients must survive long enough to be observed and included, favoring whichever group has a longer required observation window; a related risk is measurement mismatch, since in oncology trial tumor assessments follow the Response Evaluation Criteria in Solid Tumors (RECIST) while the same scenario in routine practice is typically assessed using other, less standardized criteria. A further risk is secular trend bias: outcomes can differ between a historical external-control group and a concurrent trial arm purely as a function of when each group was treated, as standard of care and diagnostic sensitivity change over time (pmc.ncbi.nlm.nih.gov). Despite these risks, a study using IPTW to build a real-world external comparator for alectinib against ceritinib in non-small-cell lung cancer (NSCLC) reported alectinib associated with a lower risk of death, hazard ratio (HR) 0.65 (95% confidence interval, CI, 0.48 to 0.88) (pmc.ncbi.nlm.nih.gov), showing adjusted external comparisons can still produce a usable, if not definitive, effect estimate.

Validation Frameworks and Sensitivity Analysis

Because no adjustment method can prove an external control arm is unbiased, regulators and methodologists have converged on sensitivity analysis as the primary safeguard. FDA's draft guidance recommends that proposed analytical methods include a strategy for dealing with missing data, since assumptions about why data are missing can themselves be unverifiable, calls for sensitivity analyses testing the impact of plausible violations of those assumptions on primary and secondary results, and further tests for vulnerability to modeling assumptions such as proportional hazards (www.fda.gov).

Regulatory guidance calls for sensitivity analysis without specifying which analyses to run or how to interpret them together, a gap a 2026 preprint addressed with an eight-part modular framework spanning heterogeneity diagnostics, source influence, no-borrowing reference cases, effective sample size, prior sensitivity, tipping-point analysis, alternative borrowing methods, and structural model sensitivity (arxiv.org) (arxiv.org). A related tool, the E-value, is the minimum strength an unmeasured confounder's associations with exposure and outcome would need to fully explain away an observed effect (arxiv.org); methodologists recommend pre-specifying its benchmark in the statistical analysis plan before unblinding (arxiv.org). Negative-control outcomes have real limits: a peer-reviewed review found they can lack both the specificity and sensitivity to detect unmeasured confounding, and that proving a null result is, strictly, impossible (www.ncbi.nlm.nih.gov).

On reporting, a public-private consortium aligned with ICH goals developed the Structured Template and Reporting Tool for Real-World Evidence (STaRT-RWE) for planning and reporting real-world evidence studies, designed to be compatible with existing bias-assessment tools and to facilitate reproducibility and validity assessment specifically. The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) similarly states that sensitivity analyses for unmeasured confounding are critical to interpreting external control arm validity, since causal-inference methods only address confounding that was actually measured (www.ispor.org).

Table 1 below summarizes how the main control-arm design options differ on the dimensions that matter for defensibility.

Table 1. Control-Arm Design Options in Clinical Trials

Design	Data Source	Contemporaneity	Primary Bias Risks	Documented Example
Randomized concurrent control	Patients enrolled in the same trial, randomly assigned	Concurrent	Minimal systematic bias by design; residual risk only from loss to follow-up or protocol deviation	Reference standard under ICH E10's five control types
Historical control (registry or natural-history)	Prior patient cohort, often a disease registry	Non-concurrent	Secular trend bias, changing standard of care, measurement drift	Eteplirsen (DMD) matched to two natural-history registries (www.accessdata.fda.gov)
Concurrent external control (RWD-based)	EHR networks, claims data, disease-specific databases	Concurrent or near-concurrent	Confounding by indication, RECIST-versus-practice measurement mismatch, immortal time bias	Alectinib versus ceritinib IPTW comparison in NSCLC
Synthetic or virtual control (AI-generated)	Generative models trained on prior trial or RWD patient records	Simulated, no directly observed comparator patient	Model training-data selection and comparability limits; unresolved regulatory terminology as of 2026	Examined in depth in a companion report on digital twins

The table makes clear that "virtual control arm" is best understood as one point on a spectrum running from fully randomized to fully simulated, not a single well-defined method. Each step away from randomization trades statistical efficiency and patient access against a specific, named bias risk that sensitivity analysis, not adjustment alone, is expected to characterize.

Implementation Considerations and Process Changes

Sponsors face a sequence of decisions that the literature above converts into a practical checklist. First, the choice of design should follow FDA's own warning: an externally controlled design is a poor fit when the anticipated treatment effect is modest, since residual bias can be large enough to obscure or fabricate an effect of that size; it is best reserved for a large expected effect, a well-characterized natural history, and a rare or serious disease where randomization is difficult to justify ethically.

Second, sponsors should select a data source with the comparability failure modes above in mind: does the candidate registry or EHR network capture outcomes on the same schedule and with the same criteria as the planned trial. Vendor-stated scale (patient counts, site counts, regulatory-decision counts) is a starting filter, not a substitute for verifying comparability, since these figures are self-reported marketing claims rather than independently audited statistics.

Third, sponsors should prespecify, before analysis, both the adjustment method (PSM, IPTW, MAIC, or Bayesian dynamic borrowing) and the accompanying sensitivity analyses, including at minimum a missing-data sensitivity analysis, a structural-assumption sensitivity analysis, and, where feasible, an E-value or negative-control diagnostic (arxiv.org). Reporting against the STaRT-RWE template from the outset can support transparent communication of real-world evidence methods and facilitate reproducibility and validity assessment (www.ncbi.nlm.nih.gov).

Fourth, and most demanding, prospective validation, testing whether the chosen methodology can reproduce a known result before it is relied upon for a real regulatory decision, is the strongest available evidence of fitness for purpose. A 2019 to 2020 validation study built EHR-derived external control groups and tested whether they could reproduce the control-arm hazard ratio from a prior, fully randomized NSCLC trial (reference trial NCT02008227, original OS HR 0.73, 95% CI 0.62 to 0.86) (www.rtihs.org); in ten of eleven analyses conducted, the EHR-derived group produced an HR estimate similar to the original RCT (www.rtihs.org). That pattern, testing a methodology blind against a trial whose answer is already known before applying it to an open question, is a defensible template for sponsors to imitate.

As an AI and life-sciences consultancy rather than a data or software vendor, IntuitionLabs advises sponsors on evidence strategy and process design of the kind described above, drawing on strategic guidance in digital transformation, AI adoption, and technology roadmapping (intuitionlabs.ai), including where a Veeva-centered clinical or regulatory data environment (intuitionlabs.ai) must accommodate an external-control-arm workflow. It does not sell external-control-arm data or software itself.

Data Analysis and Evidence

Quantifying how often external control arms are actually used, and with what data, requires distinguishing genuinely originated figures from restated ones. Grand View Research, a market research firm, sizes the broader global virtual clinical trials market (a category that spans decentralized trial technology generally, not external control arms specifically) at USD 8.8 billion in 2023, projected to reach USD 12.9 billion by 2030 at a compound annual growth rate of 5.7% (www.grandviewresearch.com), current as of the report's most recent update; no reliably originated market-size figure specific to synthetic or external control arms alone could be independently verified during this research, and that narrower figure should be treated as an open gap rather than estimated.

Peer-reviewed bibliometric work provides a more directly relevant picture of adoption. An analysis of EMA-approved cancer drugs found that 17% of cancer drugs approved in the European Union between 2016 and 2021 used external controls, and that 37% of those cases specifically leveraged real-world data (www.ebi.ac.uk). On the FDA side, a bibliometric review of FDA accelerated approvals found that of 254 total accelerated approvals granted between 1992 and 2020, 119 (47%) were oncology indications approved using a single-arm trial (www.ebi.ac.uk). Outside oncology, a review of non-oncology first-indication approvals between 2019 and 2022 identified twenty FDA and seventeen EMA approvals built on single-arm trials, the large majority using natural-history cohorts or baseline (within-patient) controls rather than a formally constructed statistical external control arm (www.ebi.ac.uk), indicating the population of trials that could benefit from an external control arm is large relative to the number that formally construct one.

FDA's NDA/BLA study-design table reports two externally controlled trials in fiscal year (FY) 2023, zero in FY2024, and four in FY2025 (www.fda.gov); these counts cover NDA/BLA submissions in the fiscal year of FDA regulatory action and do not represent all externally controlled-trial protocols. Of ten total RWE-containing New Drug Application (NDA) and Biologics License Application (BLA) submissions to CDER in FY2025, four approvals had RWE contribute to FDA's decision, and one proceeded despite RWE not contributing (www.fda.gov). A scoping review of oncology-specific external control arm studies published between 1996 and 2026 identified only 23 qualifying studies out of 629 screened, with pooled data from previous trials the single most common data source (35%, 8 of 23 studies), followed by administrative health databases and electronic medical records (17% each) (pmc.ncbi.nlm.nih.gov).

Table 2 below summarizes the statistical methods most commonly used to construct and stress-test external control arms.

Table 2. Statistical Methods for Constructing and Validating External Control Arms

Method	Purpose	Key Strength	Key Limitation
Propensity score matching (PSM)	Match trial and external patients on observed baseline characteristics	Intuitive	Only balances measured covariates; unmeasured confounding remains
Matching-adjusted indirect comparison (MAIC)	Reweight individual patient data to match a comparator's published aggregate statistics	Usable when only summary-level comparator data exists	Can sharply reduce effective sample size and still fail to achieve exchangeability
Bayesian dynamic borrowing (power/commensurate/mixture priors)	Down-weight external data in proportion to its apparent conflict with concurrent trial data	Formally incorporates uncertainty about comparability	Requires prespecified priors that are themselves open to scrutiny
E-value / quantitative bias analysis	Quantify how strong an unmeasured confounder would need to be to explain away the result	Directly addresses the confounding regulators worry about most	Only informative if a benchmark is prespecified before unblinding
Negative control outcomes	Test for a spurious association on an outcome the treatment should not affect	Cheap, easy to prespecify	Cannot prove the absence of unmeasured confounding, only sometimes detect it
STaRT-RWE structured reporting	Standardize planning and reporting of the underlying RWE study	Sets clear expectations for transparent communication of RWE methods	A reporting standard, not itself a bias-reduction method

None of these methods substitutes for the others; the literature above treats them as complementary layers, with an adjustment method selected first and sensitivity analyses, ideally several of the eight modular types described earlier, applied afterward to characterize what the adjustment could not fix.

Case Studies and Real-World Examples

Eteplirsen and Duchenne Muscular Dystrophy: Registry-Based Historical Controls

EXONDYS 51 is approved under accelerated approval for treatment of DMD in patients with a confirmed DMD-gene mutation amenable to exon 51 skipping; continued approval may be contingent on verification of clinical benefit (FDA label). FDA's medical review of the New Drug Application records that the applicant identified two DMD patient registries, the Italian DMD Registry and the Leuven Neuromuscular Reference Center registry, as the

source of external comparator data, and that thirteen historical control patients from those registries were matched to the twelve eteplirsen-treated patients in the pivotal study (www.accessdata.fda.gov). This example uses an external comparator drawn from disease-specific registries rather than a general-purpose EHR network.

Blinatumomab: A Weighted Historical Cohort for a Rare Leukemia

FDA first approved Blincyto under accelerated approval in December 2014 for Philadelphia chromosome-negative relapsed or refractory B-cell precursor acute lymphoblastic leukemia; regular approval for that indication was granted on July 11, 2017, when the indication was expanded to include Philadelphia chromosome-positive disease (FDA). The original approval's historical comparison used a 189-patient single-arm trial and a Europe-and-United-States cohort with 694 patients with complete-remission (CR) data and 1,112 patients with overall-survival (OS) data, weighted using prognostic-factor and propensity-score methods (pmc.ncbi.nlm.nih.gov). That weighted historical cohort produced a CR rate of 24% (95% CI 20% to 27%) and a median OS of 3.3 months (95% CI 2.8 to 3.6), against which the trial's own outcomes were compared to support the finding of clinical benefit. This case shows a much larger external cohort than the eteplirsen example, reflecting how much more comparator data is available for a leukemia with wider treatment history than for an ultra-rare pediatric muscular disease.

Prospective Validation in Non-Small-Cell Lung Cancer

The NSCLC EHR-derivation study discussed above under implementation considerations is worth restating here as a case study in its own right, because it is one of the few published external-control-arm exercises designed explicitly to test the method rather than to support a live regulatory submission. By deliberately picking a trial whose randomized answer was already published, the study authors could measure reproduction error directly rather than relying on plausibility arguments, a design pattern methodologists increasingly recommend as a precondition before an external control arm is used for a genuinely undecided question.

A distinct example of natural-history data supporting rare-disease development, without functioning as a formal comparator arm, comes from spinal muscular atrophy (SMA). A National Institutes of Health (NIH) summary notes that a natural-history study on SMA biomarkers, conducted through the National Institute of Neurological Disorders and Stroke (NINDS) Network for Excellence in Neuroscience Clinical Trials (NeuroNext), provided data that helped inform the nusinersen (Spinraza) clinical development program (www.ninds.nih.gov), underscoring that natural-history data can shape a development program's design and interpretation well before it is formally used as a statistical control.

Table 3 below catalogs the real-world data sources and vendors referenced throughout this report, alongside each vendor's own stated scale and role.

Table 3. Selected Real-World Data Sources and Vendors Referenced in External Control Arm Construction

Source / Vendor	Data Type	Vendor-Stated Scale (as of September 2026)	Documented Role
Flatiron Health	Oncology electronic health record (EHR) network	More than 45 global regulatory decisions informed, per vendor (flatiron.com)	—
TriNetX	Federated EHR network across participating providers	Described by vendor as the broadest federated real-world health data network available (trinetx.com)	Federated query model for observational and comparator research

Source / Vendor	Data Type	Vendor-Stated Scale (as of September 2026)	Documented Role
Komodo Health	Linked claims and clinical event data	Over 1 trillion linked records across more than 330 million de-identified patients, per vendor (www.komodohealth.com)	Healthcare Map data foundation for research and analytics
ConcertAI	Custom-abstracted oncology real-world data	Vendor claims up to 50% reduction in patient accrual needs for supported single-arm trials (www.concertai.com)	Purpose-built external control arm construction service
Verana Health (incorporating COTA)	Curated oncology and specialty historical cohorts	Vendor describes curated historical cohorts (veranahealth.com)	Historical-cohort-based synthetic control arm construction
IQVIA	Health Technology Assessment (HTA) Accelerator and RWD analytics	Vendor-conducted analysis linking RWD-based external comparators to higher HTA success rates (www.iqvia.com)	Regulatory and reimbursement strategy support

Table 3 lists vendors by the specific, verifiable claim each makes on its own site; none of these figures has been independently audited, and IntuitionLabs, as an adjacent advisory firm rather than a data or software vendor, does not appear in this table because it does not sell a comparable product.

Implications and Future Directions

FDA's February 2023 document is draft guidance distributed for comment purposes only (FDA draft guidance).

Second, FDA's NDA/BLA study-design table reports two externally controlled trials in FY2023, zero in FY2024, and four in FY2025. These counts are limited to NDA/BLA submissions in the fiscal year when FDA took regulatory action, so they do not establish an overall adoption trend.

For sponsors, the practical implication is to treat external control arm design as a validation exercise first and a statistical exercise second: prespecify the sensitivity analyses before the adjustment method is finalized, seek prospective validation against a known result wherever a suitable reference trial exists, and budget for two distinct regulatory conversations, one with FDA under its 2023 draft guidance and one with EMA under a framework that, as of September 2026, is still being written.

Frequently Asked Questions (FAQs)

What is a virtual control arm in a clinical trial?

"Virtual control arm" is commonly used for any non-randomized comparator group built from data outside the trial, though it is sometimes reserved specifically for AI-generated synthetic patients rather than directly observed comparator patients. FDA's formal terminology is "external control arm," covering both historical and concurrent external comparators.

What is a synthetic control arm in a clinical trial?

A synthetic control arm typically refers to a comparator group constructed by statistically reweighting or, in the newest AI-driven variant, generatively simulating patient-level data to represent what an untreated or standard-of-care population would look like, rather than directly enrolling that population.

What does FDA guidance say about external control arms?

FDA's February 2023 draft, non-binding guidance says sponsors must include relevant patient-level data for treatment and external control arms in marketing applications as required under FDA regulations. It recommends that proposed analytical methods include a missing-data strategy and that sensitivity analyses evaluate plausible violations of missing-data assumptions; it also warns against the design when the expected treatment effect is modest.

How does a virtual control arm differ from a placebo control?

A placebo control is a randomized, concurrent group within the same trial, sharing the same enrollment criteria, follow-up schedule, and measurement procedures as the treatment group. A virtual or external control arm is, by definition, not randomized and not necessarily concurrent, which is why it carries the specific bias risks, confounding, secular trend, and measurement mismatch, detailed earlier in this report that a placebo control does not.

Can real-world data replace a placebo or randomized control arm entirely?

Only in narrow circumstances. FDA's NDA/BLA study-design table reports two, zero, and four externally controlled trials in FY2023 through FY2025, respectively; those counts are limited to NDA/BLA submissions in the fiscal year when FDA took regulatory action and do not represent all externally controlled-trial protocols.

How is bias in a synthetic or external control arm detected and addressed?

Primarily through prespecified sensitivity analysis rather than the adjustment method alone: missing-data sensitivity analysis, structural-assumption testing, E-value analysis for unmeasured confounding, and negative-control checks are the main tools, none of which can prove the absence of bias but each of which characterizes its plausible size.

How are external control arms validated before regulatory use?

The strongest available approach is prospective-style validation against a trial whose randomized answer is already known, as in the NSCLC study that reproduced a known hazard ratio in ten of eleven analyses using EHR-derived data.

Are historical control arms still used, or have they been replaced by real-world data methods?

Historical, registry-based controls remain in active use, particularly for ultra-rare diseases where a general-purpose EHR network lacks sufficient patients, as in the eteplirsen example above; RWD and AI methods have supplemented, not displaced, them for indications with larger comparator populations.

Conclusion

External control arms, whether built from disease registries, EHR networks, prior trial data, or, in the newest and least validated variant, AI-generated synthetic patients, occupy a narrow but genuine niche in modern drug development: rare diseases, large expected effect sizes, and settings where randomization is difficult to justify ethically or practically. The regulatory record shows real momentum, FDA's 2023 draft guidance, EMA's 2024 single-arm-trial reflection paper, and EMA's 2026 concept paper committing to future external-control guidance, but submission volume remains small and uneven rather than rapidly scaling. The methodological literature treats exchangeability as an ideal that adjustment methods approximate rather than achieve, and sensitivity analysis, not the adjustment method itself, as the primary safeguard against overstating what an external comparison can show. Readers evaluating a specific AI-driven synthetic control platform should consult the companion IntuitionLabs report on digital twins for the technology-specific detail this report deliberately leaves to that treatment.
