Synthetic Data in Pharma: A Guide to Acceptance Criteria

By InuitionLabs.ai • 10/13/2025 • 35 min read

synthetic data | pharmacovigilance | clinical trials | acceptance criteria | data fidelity | data privacy | generative ai | pharmaceutical research

Executive Summary

Synthetic data are increasingly seen as a transformative solution to the data gaps and privacy constraints in pharmaceutical research, including pharmacovigilance (PV) and clinical development (pmc.ncbi.nlm.nih.gov) (jamanetwork.com). By replicating the statistical structure and relationships of real patient data without disclosing individual identities, synthetic datasets promise to enable wider data sharing, augment under-represented populations, and accelerate AI-driven analytics, while *ideally* satisfying privacy regulations (pmc.ncbi.nlm.nih.gov) (www.frontiersin.org). However, for synthetic data to "survive inspection" – that is, to be accepted by regulators, ethicists, and practitioners – they must meet stringent criteria. These include quantitatively preserving key distributions and predictive relationships (data **fidelity**), enabling valid analyses (data **utility**), and provably protecting individual privacy (data **privacy**) (jamanetwork.com) (journals.plos.org). Current frameworks highlight trade-offs: data closer to reality yield more utility but pose greater re-identification risk (journals.plos.org) (jamanetwork.com).

This report provides a deeply detailed examination of synthetic data in pharmaceutical contexts, with a focus on pharmacovigilance and clinical research. It covers historical background, definitions and types of synthetic data, relevant regulatory and ethical considerations, technical generation methods, and rigorous validation strategies. We survey multiple perspectives, including academic reviews, regulatory commentary, and practical case studies, providing concrete data on performance and constraints. Through numerous examples – from synthetic electronic health records to large-scale simulated genomic cohorts – we analyze how synthetic data have been validated (or found lacking) and propose best practices for acceptance. A particular emphasis is placed on "acceptance criteria": quantifiable metrics and qualitative standards by which synthetic datasets may be judged fit for PV signal detection or clinical trial simulation. Where applicable, we incorporate expert opinions and current research findings (e.g., GAN-based hematology datasets (ascopubs.org) (ascopubs.org), synthetic claims records (pubmed.ncbi.nlm.nih.gov), or synthetic risk models (www.frontiersin.org) (www.frontiersin.org)) to ground the discussion. We also discuss ongoing challenges – from statistical biases to regulatory uncertainty – and emerging solutions (such as differential privacy and evaluation frameworks). Finally, future directions for policy and technology development are explored, aiming to guide stakeholders in harnessing synthetic data responsibly.

Introduction

Definitions and Background

Synthetic data broadly refers to *"data that have been created artificially (e.g., through statistical modeling [or] computer simulation) so that new values and/or data elements are generated"*, designed to emulate the structure and relationships of actual patient data without containing any real individual's information (pmc.ncbi.nlm.nih.gov). By this definition (aligned with the FDA's glossary for digital health/AI), synthetic records reproduce statistical properties of real datasets while replacing sensitive values with generated ones (pmc.ncbi.nlm.nih.gov) (www.frontiersin.org). The U.S. Census Bureau

similarly defines synthetic data as output of statistical models that capture real data's multivariate distribution (journals.plos.org) (pmc.ncbi.nlm.nih.gov). Importantly, synthetic data **do not include any actual patient records**, distinguishing them from true or anonymized data (pmc.ncbi.nlm.nih.gov) (jamanetwork.com).

Synthetic datasets can be categorized into **fully synthetic**, **partially synthetic**, or **hybrid** approaches (journals.plos.org) (journals.plos.org). *Fully synthetic data* replace all original records with entirely generated values, yielding very low re-identification risk but potentially lower analytical fidelity (journals.plos.org). *Partially synthetic data* substitute only sensitive variables (e.g. names, addresses, some attributes), retaining much of the original structure but carrying higher disclosure risk (journals.plos.org). *Hybrid data* mix real and synthetic elements – for example, pairing each real record with a nearby synthetic one – which aims to balance privacy and utility (journals.plos.org). Different degrees of "syntheticness" exist along a spectrum (from synthetic structural data with no analytics value to near-identical data with high utility) (journals.plos.org) (journals.plos.org). Table 1 (below) summarizes these types and their trade-offs.

Type of Synthetic Data	Description	Privacy Risk	Analytical Utility	Source
Fully Synthetic	Every record and field is artificially generated (no real data included).	Very low (no real patient information present).	Can reproduce global statistics, but may miss subtle real-world correlations; typically lower fidelity for complex analyses (journals.plos.org).	PLOS 2023 (journals.plos.org)
Partially Synthetic	Sensitive attributes (identifiers or </current_article_content>rare values) are replaced with synthetic values; non-sensitive data remain real.	Moderate (some original data remain, higher re-ID risk than fully synthetic) (journals.plos.org).	Higher data utility than fully synthetic (retains real data distribution for non-sensitive fields) (journals.plos.org).	PLOS 2023 (journals.plos.org)
Hybrid	Combines real and synthetic, e.g. merging each real record with a corresponding generated record.	Moderate (because real data are still present, but distorted) (journals.plos.org).	Generally highest utility for analysis (preserves joint distributions) (journals.plos.org).	PLOS 2023 (journals.plos.org)

(Table entries extracted and paraphrased from Gonzales et al. 2023 (journals.plos.org) (journals.plos.org).)

Synthetic data generation originated in the statistics community (e.g. Donald Rubin's synthetic dataset for the U.S. Census in the 1990s) as a way to share data safely (www.pnas.org). In healthcare, synthetic data research emerged amidst growing data sharing and privacy concerns. Recently, interest has been reignited by advances in AI (notably deep learning) that can generate high-dimensional data on demand (www.pnas.org) (www.frontiersin.org). For example, generative adversarial networks (GANs) – powerful AI models with a "generator" and "discriminator" – have been used to create synthetic EHR records that are statistically indistinguishable from real ones (www.frontiersin.org). Table 2 (below) highlights representative projects leveraging synthetic data in healthcare, from electronic health records to genomic cohorts, and reports how closely they matched real data in specific measures.

Table 2: Examples of Synthetic Data Projects in Healthcare/Pharma

Use-Case / Data Domain	Approach / Tool	Validation Measure	Findings / Performance	Ref.
Massachusetts Synthetic EHR (Synthea)	Rule-based, public health guidelines simulator (Synthea)	Clinical quality indicators (e.g. cancer screening rate, mortality)	Reproduced demographics and care probabilities well, but significantly underestimated adverse outcomes (e.g. 0.7% vs 7–8% mortality after COPD exacerbation) (bmcmedinformdecismak.biomedcentral.com); did *not* model complications or BP control realistically (bmcmedinformdecismak.biomedcentral.com).	Wuri et al. 2019 (bmcmedinformdecismak.biomedcentral.com)
Synthetic Claims Data (Manitoba, Canada)	Statistical simulator (OSIM2) and enhanced model (ModOSIM)	Concordance coefficient for drug utilization metrics	ModOSIM achieved ~88% concordance with real registry on average prescription-days, whereas the simpler OSIM2 was only ~16% (pubmed.ncbi.nlm.nih.gov). Partially replicated sex/age distributions, but both models under/over-estimated numbers of medications per person (pubmed.ncbi.nlm.nih.gov).	Gronsbell et al. 2023 (pubmed.ncbi.nlm.nih.gov)
Heart Failure EHR (26k patients)	Deep learning GAN (Tabular GAN) for EHR synthesis	Predictive model AUC on synthetic vs real data	Trained ML (DNN) on fully synthetic EHR records achieved AUC=0.80 predicting 1-year mortality (vs ~0.80 with real training) (www.frontiersin.org). Synthetic data were "statistically indistinguishable" from real EHR, enabling ML analysis while eliminating PHI (www.frontiersin.org) (www.frontiersin.org).	Guo et al. 2020 (www.frontiersin.org) (www.frontiersin.org)
Hematology Registered Data (7,133 pts)	Conditional Wasserstein GAN (tabular GAN)	Composite fidelity score (CSF/GSF), privacy metrics (NNDR)	Generated synthetic cohorts with *"high fidelity"* to clinical/genomic variables (ascopubs.org). Fidelity metrics exceeded defined thresholds (≥85% agreement) and privacy metrics (nearest-neighbor ratios 0.6–0.85) were satisfied (ascopubs.org) (ascopubs.org). Synthetic augmentation captured trial endpoints and enabled discovery of genomic associations consistent with larger real cohorts (ascopubs.org) (ascopubs.org).	D'Amico et al. 2023 (ascopubs.org) (ascopubs.org)
Genomic Cohort (1M genotypes)	Reference-based resampling (HAPNEST)	Allele frequency (MAF), LD patterns, population structure, PRS	Produced a 1,008,000-person synthetic biobank with 6.8M SNPs that preserved key statistics of real 1000Genomes data (pmc.ncbi.nlm.nih.gov). LD and ancestry distributions closely matched true data, enabling PRS comparisons across populations (pmc.ncbi.nlm.nih.gov). Computationally efficient: 20k-SNP genome ~15 min (1 thread) (pmc.ncbi.nlm.nih.gov).	Wharrie et al. 2023 (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov)

These examples demonstrate both the potential and pitfalls of synthetic data. The Synthea project shows that synthetic patient cohorts can reliably mimic demographic and service-utilization patterns, but may fail to capture deviations from guidelines or post-treatment outcomes (bmcmedinformdecismak.biomedcentral.com). Conversely, advanced methods like GANs have produced synthetic datasets that successfully train predictive models comparably to real data (www.frontiersin.org) (www.frontiersin.org). However, nothing should be taken for granted: even state-of-the-art approaches must be rigorously validated. The remainder of this report examines how such validation is (or should be) performed, and what metrics determine acceptability in PV and clinical contexts.

Synthetic Data in Pharmacovigilance and Safety Surveillance

Pharmacovigilance (PV) – the monitoring of adverse drug reactions (ADRs) and safety signals – is inherently data-intensive. Traditional PV relies on voluntary reports, structured databases (like FAERS/VigiBase), and increasingly on real-world sources (EHRs, claims, social media) (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). However, patient privacy rules and data fragmentation often limit access to large, labeled datasets for training PV algorithms (such as NLP for extracting ADRs from clinical notes). Synthetic data have been proposed as one solution to these challenges. By generating *realistic but fake* adverse event (AE) datasets, companies can train AI systems without handling identifiable records. For example, given a data-scarce safety review scenario, synthetic case reports could be generated (e.g. via text GANs) that preserve the narrative patterns of real reports, enabling algorithm development without privacy risk [**citation needed**].

Beyond NLP, synthetic PV data might include simulated electronic healthcare event timelines or "digital twins" of patient populations at risk. Such data could help test signal-detection algorithms under controlled conditions: known safety signals can be planted in a synthetic dataset to probe detection performance. In principle, one could also simulate entire drug safety databases (for example, a synthetic VigiBase) to benchmark new disproportionality or machine-learning methods without violating confidentiality (www.pnas.org) (jamanetwork.com).

However, the literature provides few full case studies of synthetic data in PV, reflecting that the field is just beginning to explore this application. A 2022 survey of AI in PV noted the importance of data quality and validation (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov), and similar concerns apply to synthetics. **Acceptance Criteria:** Synthetic PV data must maintain the true distributions and relationships between drugs, events, and covariates. For example, if a real-world dataset shows a strong association (e.g. disproportionality) between Drug A and Adverse Event X, a valid synthetic dataset should preserve that association—and not introduce spurious correlations. Key criteria would include:

- **Statistical Fidelity:** Marginal and joint distributions of drug usage, co-medications, and event types should closely match those in benchmark data (when available) (journals.plos.org). If synthetic AE counts by age/sex deviate significantly from expected rates, findings could mislead PV analysts.
- **Signal Preservation:** Known safety signals (either established or newly inserted) should remain detectable. If a synthetic dataset was used to simulate a known adverse reaction, algorithms should flag it similarly as in real data.
- **Covariate Balance:** If modeling risk factors (e.g. comorbid conditions, genetic profiles), synthetic data should maintain realistic covariate distributions so that methods like propensity scoring or regression models trained on synthetic data generalize to real-world deployments.

Because PV often deals with rare events, an additional concern is **sampling realism**. For very rare ADRs, synthetic oversampling (analogous to SMOTE) might be used to enhance dataset balance (www.citeline.com). Yet overdoing this could distort incidence rates. Thus, any synthetic augmentation for PV must be carefully parameterized and validated for epidemiological plausibility.

Finally, just as for any patient data, synthetic PV datasets must have *no re-identification risk*. Spurious preservation of unique combinations of attributes (especially for rare conditions) could inadvertently re-create patient identities (journals.plos.org) (jamanetwork.com). We return to privacy validation in Section 5 below.

Synthetic Data in Clinical Research and Trials

Synthetic data are similarly finding traction in clinical development. Two major domains are: (1) **Clinical trial design and analysis**, and (2) **Digital health/device prototyping and algorithms**.

Synthetic Control Arms and Trial Simulation

Perhaps the most talked-about application is the **synthetic control arm**. In a trial with a single treatment arm, a synthetic dataset of patients who "received placebo" could theoretically serve as a comparator. Some call this an *external control*, but true synthetic control arms (generated by models) have not yet been accepted in regulatory submissions (www.pnas.org). In a recent perspective, a UK regulator noted that "no one has used a truly synthetic cohort as a control group in an approved trial" (www.pnas.org). However, the concept is compelling: by simulating patient trajectories under standard of care (or placebo), one could reduce trial size or avoid randomizing patients to inferior treatments (especially in rare diseases). Preliminary discussions (e.g. U.S. FDA working groups) suggest regulators may eventually allow this for planning or hypothesis-generation, but only as *supportive* evidence, not a sole basis for approval (www.pnas.org).

Even if full synthetic arms are not yet sanctioned, related in silico methods do see use. For example, physiologically-based PK/PD models routinely generate virtual patient data (dose–response curves, biomarker levels, etc.), which regulators accept for drug development (pmc.ncbi.nlm.nih.gov). This is technically a form of synthetic data (process-driven), deeply rooted in pharmacometrics. Similarly, advanced quantitative systems pharmacology (QSP) models can simulate disease progression under new therapies. These traditional mechanistic simulations can be seen as the "process-driven" end of the synthetic data spectrum (pmc.ncbi.nlm.nih.gov). They have established validation standards (e.g. matching known PK parameters) and are generally trusted by regulators because they are fully transparent models of biology.

By contrast, purely **data-driven synthetic clinical trials** (e.g. using GANs on limited patient data) are nascent. A key acceptance criterion here is **outcome concordance**: does the synthetic trial yield the same efficacy/safety results as a real trial? For instance, D'Amico et al. demonstrated that synthetically augmenting 187 trial patients with a GAN could *almost exactly* recapitulate the trial's clinical endpoints (ascopubs.org). Future guidelines would likely require demonstration that key endpoints (e.g. survival curves, response rates) computed on synthetic control groups closely match those on real arms, before trusting the synthetic version.

In general, synthetic or model-based approaches in clinical trials must ensure:

- **Regulatory comparability:** Synthetic cohorts should be statistically indistinguishable, by predefined criteria, from conventional external controls (historical or concurrent RWD). Regulators

will scrutinize definitions (as CPT & EMA emphasize the need for consistent language around "external controls" vs. "synthetic arms") (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov).

- **Reproducibility:** Independent researchers or reviewers should be able to reproduce the synthetic data generation process (i.e. full transparency of algorithms and parameters). This transparency is crucial for "surviving inspection" by regulators and ethics boards.
- **Model Validation:** If synthetic data come from ML (e.g. GANs), the underlying model must be validated against training data (potentially using reserved real data) to show it generalizes correctly. For example, in D'Amico's work, the team reserved some real samples to confirm the GAN was not simply memorizing the input (ascopubs.org).
- **Safety Margins:** Synthetic datasets used to simulate trials must never replace initial human studies. The consensus is that *eventually* synthetic or in silico results would only supplement – not supplant – empirical clinical evidence (www.pnas.org). Regulators anticipate synthetic models to come into play most in early design or in generating hypotheses for under-studied populations (e.g. pregnant women) (www.pnas.org).

Other Clinical and Development Uses

Beyond trials, synthetic data can aid clinical research in other ways. For example:

- **RWD Augmentation:** Pharmaceutical companies often analyze real-world data (claims, EHR) for safety and outcomes studies. Synthetic data can *fill gaps* when patient numbers are too low. Citeline noted synthetic data's use in augmenting rare disease cohorts to improve statistical power (www.citeline.com). Such augmentation might, for example, double the size of a pediatric epilepsy registry while preserving epidemiological trends.
- **Algorithm Validation:** Medical AI models (e.g. diagnostic image classifiers) can be tested on synthetic datasets to simulate worst-case scenarios or unusual conditions that lack real samples. If an AI model fails on synthetic rare cases, it signals a need for caution.
- **Education and Systems Testing:** Synthetic EHRs are already used to train clinicians and to validate health IT systems. A large synthetic database can stress-test a new data-cleansing pipeline or train students in clinical informatics without risking PHI. Though not a regulatory acceptance issue, it illustrates a practical acceptance: a dataset survives "inspection" if it faithfully exercises the system under realistic conditions.

In all clinical applications, key acceptance criteria overlap with those for PV: *realism* of patient features and outcomes, *reproducibility* of insights, and demonstrable *no risk* to actual patients' privacy. As one expert noted, synthetic data have "the potential to be used to generate larger sample sizes and so increase the statistical power of analyses" (www.pnas.org) – but only if the synthetic augmentation truly mirrors complex patient characteristics.

Methods of Synthetic Data Generation

Synthetic data can be produced by a variety of computational methods. We divide these broadly into **process-driven models** and **data-driven (statistical or AI) models** (pmc.ncbi.nlm.nih.gov).

- **Process-Driven Models:** These rely on mechanistic understanding of the system. In pharmacology, this includes models like physiologically-based pharmacokinetics (PBPK), physiologically-based pharmacodynamic (PBPD) models, or more general systems biology models. For example, a PBPK model that encodes drug absorption and elimination pathways can simulate virtual patients' blood concentrations under different dosing regimens. These virtual patients define a synthetic dataset of concentration–time points. Regulatory agencies have long endorsed such simulations (e.g. FDA guidance on PBPK) (pmc.ncbi.nlm.nih.gov). Similarly, in-silico trials that use biophysically detailed models (e.g. finite-element heart models) generate synthetic MRI or ECG data that "look" and behave like real ones. **Acceptance:** Process-driven synthetic data often come with rigorous mathematical transparency, so regulators can inspect the equations and assumptions. The usual validation approach is to check that the model output matches historical observed data (face validity) – for instance, ensuring a PBPK model predicts known drug levels across populations.

- **Statistical/Data-Driven Models:** These use algorithms to **learn** data distributions and then sample new records. Key techniques include:

- **Probabilistic Imputation:** Simple methods like multiple imputation or bootstrapping from distributions. For example, one might fit a Bayesian network or a joint probability table to existing data and draw samples. These methods are often used by statisticians for tabular data [8].

- **Generative Adversarial Networks (GANs):** Deep learning models where a "generator" network produces fake data and a "discriminator" network tries to distinguish fake from real, refining the generator over many iterations. GANs have been applied to images, texts, and *tabular* health data (www.frontiersin.org) (ascopubs.org). For instance, modern *tabular GAN* architectures (e.g. CTGAN, TVAE, WiGAN) can handle mixed numeric/categorical data. In D'Amico's hematology study, they used a *conditional Wasserstein GAN with gradient penalty*, specifically tailored to clinical tabular data (ascopubs.org). This approach enabled modeling complex correlations between labs, genomics, and outcomes.

- **Variational Autoencoders (VAEs):** These feed data through an "encoder" to a latent distribution and then decode to reconstruct input. By sampling from the latent space, VAEs can generate new synthetic examples. They typically yield more "smooth" but at times blurrier synthetic records than GANs.

- **Diffusion Models:** A newer class (popular in images) that gradually denoises random noise into structured outputs. These have **not** yet seen wide adoption for typical clinical tabular data but could be relevant for image or waveform synthesis in the future.

Notably, many synthetic data generators are **"black box"** AI models that lack interpretability. This raises special concerns in pharma: unlike in consumer apps, safety and efficacy demand fully understood methods. Nonetheless, the rapid progress in ML has meant that techniques like GAN/VAEs are increasingly accessible (with tools like PyTorch and TensorFlow implementations).

Toolkits: There are both commercial and open-source tools. Examples include:

- **Synthea:** An open-source rule-based patient generator (used in Table 2 above) (bmcmedinformdecismak.biomedcentral.com).
- **MDClone (Civica):** A commercial platform that claims to produce high-fidelity synthetic EHR data via a "data cloning" approach. While details are proprietary, published experiences suggest it indexes real data and draws samples that statistically match correlations.
- **HAPNEST:** An open-source tool (see Table 2) specifically for genetic data (pmc.ncbi.nlm.nih.gov).
- **CTGAN (Synthetic)** and related libraries in Python that implement GAN variants for tabular data.

Regardless of the model, all synthetic generation methods share a need for **training data**: some real dataset (or summary statistics) to base the synthetic creation on. Hence, a critical acceptance criterion is simply that *the synthetic process* was indeed derived from an appropriate real data source. If a clinical trial sponsor submits synthetic controls, regulators will ask: *What original data were used to create these synthetic patients?* The process should be fully documented, with any randomness seeded or controlled (to allow repeatability).

Validation, Quality Metrics, and Acceptance Criteria

A synthetic dataset is only useful if it faithfully supports the analyses for which it is used. Thus, rigorous **validation and evaluation** are essential. Unfortunately, as of today there is no universally accepted "synthetic data standard" or single metric (jamanetwork.com) (journals.plos.org). Instead, experts recommend multi-faceted assessment along three axes: **fidelity**, **utility**, and **privacy risk** (jamanetwork.com). These correspond closely to the concerns of regulators and end-users:

- **Statistical Fidelity:** Measures how closely the synthetic data replicate the statistical properties of real data. Typical checks include comparing means, variances, and higher moments of variables, and more generally the joint distributions. For example, we might run Kolmogorov–Smirnov tests on continuous fields, or chi-square tests on categorical distributions. Metrics such as *Wasserstein distance* or *Maximum Mean Discrepancy (MMD)* have been used to quantify global distribution differences. In some studies, investigators compute fidelity scores or correlation coefficients across many features. The D'Amico et al. hematology study computed a **clinical synthetic fidelity (CSF)** and **genomic synthetic fidelity (GSF)** by averaging multiple metrics (ascopubs.org), requiring them to exceed a threshold (85%) to consider the synthetic set acceptable. In practice, regulators would likely demand that critical endpoints or covariate distributions match within pre-specified tolerances (e.g. *no significant shift in disease incidence by age group*).
- **Preservation of Operational Utility:** Even if two datasets have matching univariate stats, the real test is whether they lead to **equivalent analysis results**. For example, if one fits a risk-prediction model or computes a treatment effect on synthetic data, the parameter estimates and confidence intervals should agree closely with those from real data. This is essentially *outcome validation*. Some authors argue that the highest bar is to insist synthetic data reproduces the answers of research (e.g. with no "different answer") (www.pnas.org). For instance, Guo et al. showed that models trained on fully synthetic EHR predicted patient mortality with virtually the same accuracy (AUC) as models trained on real data (www.frontiersin.org). Any acceptable synthetic dataset should be tested in this fashion for its intended use-case. For PV, that might mean comparing signal-detection metrics; for trials, it means comparing effect sizes or survival curves.
- **Privacy and Disclosure Risk:** The synthetic data **must not allow re-identification of real individuals**. This implies, for one, that there should be no exact copies of any record from the real source (beyond trivial synthetic "noise" around them). Authors often measure the *"identical match share" (IMS)* or compute the distance of each synthetic record to its nearest real neighbor. A well-regarded rule-of-thumb is that the *nearest-neighbor distance ratio* (NNDR: ratio of distance to first vs second nearest real neighbor) should be neither too close to 0 (too similar) nor exactly 1

(dissimilar). D'Amico et al., for example, recommended an NNDR in the range **0.6–0.85** (ascopubs.org). More generally, if we consider differential privacy as a metric, most advise an epsilon ≤1 for high privacy, though achieving this for high-dimensional clinical data is challenging. Additional risk comes from adversarial attacks (e.g. membership inference), so synthetic data often incorporate noise or privacy mechanisms to mitigate it (journals.plos.org) (jamanetwork.com). Crucially, no acceptable synthetic dataset is truly "risk-free" – it is regarded as disclosure risk *reduced* rather than *eliminated* (jamanetwork.com). Therefore, *risk evaluation* (like attempt at re-identification by an attacker) is part of the acceptance process.

In summary, **synthetic data should be validated on multiple fronts**. Practical acceptance criteria often boil down to *"close enough"* agreement with real data on relevant measures plus *provably low* privacy risk. Table 3 below outlines key metrics and their targets as reported by recent studies.

Metric	What It Measures	Target/Threshold	Source / Notes
Fidelity Scores (e.g. CSF/GSF)	Composite score averaging statistical similarity across many features (e.g. means, variances, correlations).	High fidelity; in one study CSF/GSF ≥85% was required (ascopubs.org).	Synthetic data analyses should yield similar descriptive stats and model parameters as real data.
Nearest-Neighbor Ratio (NNDR)	Ratio of distances from each synthetic point to its nearest and second-nearest real neighbors.	Typically **0.6–0.85** recommended (ascopubs.org) (0.5=synthetics are too dissimilar; 1.0=too identical).	Ensures synthetic points are neither verbatim copies nor wholly implausible outliers.
Exact-Match Incidence	Fraction of synthetic records identical to any original record (zero distance).	Essentially **0%** (no exact duplicates allowed) (ascopubs.org).	Even a few exact matches are unacceptable for privacy.
Distributional Tests (e.g. KS, MMD)	Statistical tests comparing real vs synthetic variable distributions (continuous or categorical).	Non-significant differences (e.g. p>0.05 in KS, or low MMD) on key variables.	Used for initial sanity check of individual fields and multivariate dependencies.
Predictive Performance	Compare ML model metrics (e.g. AUC, coverage) when trained on synthetic vs real data.	Minimal drop in performance. E.g. AUC synthetic ≈ AUC real (www.frontiersin.org).	Ultimately, downstream analytics should agree within a few percentage points, at least.
Re-identification Risk (e.g. DP ε)	Quantifies privacy leakage (e.g. differential privacy epsilon).	$\varepsilon \leq 1$ (often cited for strong privacy, though few datasets reach this).	Smaller ε means stronger privacy guarantee. Currently difficult for rich healthcare data.
Utility for Intended Use	Qualitative check (e.g. clinical studies re-run).	Synthetic analyses should yield same conclusions as on real data (no "different answer") (www.pnas.org).	The ultimate litmus test: would a decision based on synthetic data match one based on real data?

(Table based on strategies discussed in the literature (ascopubs.org) (jamanetwork.com) (www.pnas.org).)

Notably, **no single score suffices**. The JAMA Forum and PLOS digital health reviews emphasize that existing metrics are insufficient alone (jamanetwork.com) (journals.plos.org). Regulators will likely require multiple, overlapping validations: statistical tests, expert review of plausibility, and transparency of methods. In the CVR (critical limitations) perspective, one must provide not just numbers but reasoning for "why synthetic data is justified, and any assumptions or caveats" – for example, documenting if rare outliers were suppressed for privacy.

As a compelling example, Foraker et al. (Washington U.) compared analyses on synthetic vs original EHR and reported "not gotten a different answer" so far (www.pnas.org). While promising, such anecdotal reports need systematic backing. In practice, before accepting synthetic data for PV or clinical use,

sponsors would present a validation dossier: comparing patient demographics, outcome rates, and analytical results with those from the source data (or a high-quality proxy).

Regulatory and Ethical Considerations

Synthetic data in healthcare intersects with data protection laws and regulatory guidelines. There are no regulations written *specifically* for synthetic data used in drug development or PV. Instead, synthetic data must navigate existing frameworks:

- **Data Protection Law:** In the EU's GDPR, synthetic data generated from personal data is considered **processing of personal data** (jamanetwork.com). This means that, despite having no identifiable real records, synthetic datasets are not automatically "free" of GDPR oversight. Whether synthetic data are considered anonymous (nonpersonal) is debated. The EU AI Act (effective Aug 2024) explicitly mentions synthetic data as a privacy-preserving alternative in high-risk AI systems (jamanetwork.com), but it does not define privacy thresholds for synthetic data. In practice, synthetic data are treated similarly to anonymized data – requiring assurance that generative models did not simply memorize individuals (jamanetwork.com) (jamanetwork.com). The JAMA viewpoint states clearly: *"Creating synthetic data from real personal data is considered processing under the EU's GDPR… and whether synthetic data remain personal is a complex issue"* (jamanetwork.com). Thus, any use of synthetic healthcare data still demands Data Protection Impact Assessments and safeguards (e.g. encryption, limited access) as if handling personal data.

- **Informed Consent and Secondary Use:** For clinical trial or registry data repurposed in synthetic generation, patient consent may need to cover this use. Some ethical boards might view synthetic data generation as an additional use not anticipated in the original consent. However, because synthetic data contain no identifiable patients, some argue it is low-risk. Best practice is to include mention of possible synthetic derivations in initial study protocols or data usage policies.

- **Regulatory Guidance:** Until regulators issue formal guidelines, practice is guided by general principles. Notably, the international "Guardrails for Synthetic Control Arms" and similar position papers urge clarity of definitions. The CPT PSP review (Pasculli et al.) notes that FDA and EMA define *external controls* as any data "from another setting", but they do **not** address generative synthetic data explicitly (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). In the UK, MHRA informally recognizes synthetic data as "artificial data that mimic the properties of real data" (without specifying acceptance criteria) (pmc.ncbi.nlm.nih.gov). In short, researchers should assume that synthetic datasets used for any regulatory purpose will be subject to the same scrutiny as real data: requiring documentation of how data were generated, validated, and governed.

Ethically, synthetic data promise to **reduce harm** (no patient privacy lost) and **increase fairness** (by enabling data for underrepresented groups) (jamanetwork.com). However, they can also *exacerbate* biases if not carefully controlled. The JAMA article warns that synthetic data may unintentionally encode and amplify biases present in the source data (jamanetwork.com) – a critical concern in PV and patient safety decisions. Thus, any synthetic data project must also consider bias mitigation (e.g. re-weight or post-process synthetic outputs to correct known imbalances).

Case Studies and Applications

To illustrate these principles in action, we review several detailed case studies where synthetic data have been used or evaluated in pharmaceutical/healthcare contexts. Each demonstrates particular

acceptance issues.

- **PLOS Quality Metrics (Synthea Massachusetts).** Wuri et al. (2019) tested an open-source synthetic patient generator (Synthea) by comparing clinical quality metrics (screening rates, mortality, complication rates) calculated on synthetic vs real data (bmcmedinformdecismak.biomedcentral.com). Their findings were mixed: while basic demographics and service offers matched Massachusetts stats, outcome-based measures did not. For example, synthetic physiology yielded only 63% colorectal cancer screening vs 77% in real data, and 0.7% COPD mortality vs 7-8% actual (bmcmedinformdecismak.biomedcentral.com). The study concluded that synthetic data must incorporate care variation (quality measures) into their logic to be truly realistic (bmcmedinformdecismak.biomedcentral.com). **Acceptance insight:** Validation against domain-specific performance metrics (here, healthcare quality scores) flagged limitations. The acceptance criterion fails if synthetic patients have systematically **healthier outcomes** than real ones, as these do in Synthea.

- **Regulatory-RWD Simulation (OSIM and ModOSIM).** In a Canadian study, synthetic administrative health records were generated via two methods: the off-the-shelf OSIM2 simulator and an improved "ModOSIM" algorithm (pubmed.ncbi.nlm.nih.gov). The real dataset contained 169 million drug dispenses; they simulated 1M-person synthetic sets. While demographics (age/sex) matched well, OSIM2 greatly under-estimated drug counts and durations (only ~16% concordance with real data), whereas ModOSIM reached ~88% concordance (pubmed.ncbi.nlm.nih.gov). In other words, only the enhanced modeling produced data **similar enough** to real usage patterns. **Acceptance insight:** This highlights that simple generative rules can fail on complex data. A synthetic dataset might look plausible at first glance, but detailed comparison (drug days, polypharmacy rates) revealed OSIM2 was unacceptable for analytic use. Acceptable synthetic PV/clinical data may thus require iterative refinement: try a candidate generator, compute concordance or distance metrics as above, and tune until acceptable (much as ModOSIM outperforming OSIM2).

- **Synthetic EHR for Predictive Modeling.** The Washington Univ. team (Guo et al.) generated synthetic EHR data for 26,575 heart failure patients using a tabular GAN (www.frontiersin.org). They demonstrated that machine learning models built on the synthetic cohort performed equivalently to real-data models: the deep neural network had AUC≈0.80 on synthetic vs 0.80 on real data (predicting 1-year mortality) (www.frontiersin.org). Critically, they found that "synthetic EHR data are statistically indistinguishable from the original… and can be analyzed as if they were original data" (www.frontiersin.org). They also noted practical advantages: using synthetic data "eliminates risk of exposure of EHR data, speeds time-to-insight, and facilitates data sharing" (www.frontiersin.org). **Acceptance insight:** Here, the acceptance criterion was that a clinically meaningful outcome (risk of death) was predicted equally well. The authors also ensured no synthetic patient re-used an actual PHI entry, eliminating privacy risk. This case shows that, for certain tasks, synthetic data can be directly substituted for real data without loss of analytic performance (www.frontiersin.org).

- **Hematology Clinical-Genomic Data (GAN by D'Amico et al.)** This 2023 study is exemplary because it was specifically designed to meet "acceptance by construction." The authors collected comprehensive clinical and genomic data on 7,133 patients with myelodysplastic syndromes or AML. They trained a conditional GAN (with special architectural refinements) to generate synthetic patients. Crucially, they developed a *Synthetic Validation Framework*: they defined numerical metrics for data fidelity and privacy, and set passing thresholds (ascopubs.org) (ascopubs.org). For synthetic cohort evaluation, they reported that *all* metrics were satisfactorily met: their synthetic data had CSF/GSF fidelity scores above 85%, and their nearest-neighbor privacy metrics fell in the safe range (ascopubs.org) (ascopubs.org). They then demonstrated utility: starting from only 944 real MDS cases, they generated 3× as many synthetic patients and were able to identify genetic risk groupings that eventually matched those from a much larger real dataset of 2043 patients (discovered years later) (ascopubs.org). Moreover, they generated synthetic control arms for a luspatercept trial (N=187) that *recapitulated the actual trial endpoints*. Finally, they even built a public website so other clinicians could generate synthetic cohorts from registry data. **Acceptance insight:** This work sets a gold-standard example. They defined clear acceptance criteria (fidelity ≥85%, no excessive memorization), documented the model, and showed synthetic-driven research leading to real discoveries. Regulators would be impressed by such thorough validation and transparency. It shows synthetic data can *survive regulatory inspection* if one specifies and meets numeric benchmarks in advance (ascopubs.org) (ascopubs.org).

These real-world studies highlight that the bar for "surviving inspection" is tangible. A synthetic dataset may look "plausible" on cursory review, but only detailed measurements and comparisons will reveal its weaknesses or strengths. Stakeholders requiring acceptance (e.g. FDA, IRBs, data protection officers) will want evidence: e.g. tables or plots comparing distributions, ML performance, and disclosure metrics for the synthetic vs original data.

Challenges and Limitations

Despite the promise, synthetic data come with important caveats. Notably:

- **Data Leakage and Memorization:** Generative models can inadvertently memorize and reproduce training data verbatim (jamanetwork.com) (jamanetwork.com). Famous incidents (such as GPT-3 leaking phone numbers) underscore that even well-trained networks can overfit rare patterns (jamanetwork.com). In healthcare, this means that very rare patient profiles (e.g. a combination of diagnoses) might be unintentionally replicated, risking confidentiality breaches. **Mitigation:** Authors recommend techniques like differential privacy (adding calibrated noise during training) and post-hoc checks (e.g. measuring the distance of synthetic examples from nearest real records) (journals.plos.org) (ascopubs.org). In practice, any synthetic pharma dataset must be scanned for unique outliers and dropped or perturbed if too close to originals.

- **Bias Amplification:** Synthetic data generation inherently learns from existing data. This means any biases or artifacts in the source will likely propagate into the synthetic output (jamanetwork.com). For example, if a clinical dataset under-represents certain ethnic groups, a naive GAN may produce even fewer samples of those groups or exaggerate any spurious correlation those patients had. The JAMA article warns that synthetic data *"may perpetuate or even amplify unresolved biases and spurious correlations from the original data"* (jamanetwork.com). In PV, this could mean missing signals in disadvantaged populations or erroneously inflating risk factors for over-represented groups. **Mitigation:** Careful bias analysis is needed. One approach is to enforce parity or oversample underrepresented groups during generation (targeted synthetic augmentation) (jamanetwork.com), although preserving clinical realism remains challenging. Another is to apply fairness constraints or reweighting in the synthetic-generation algorithm itself.

- **Modeling Rare Events:** By definition, synthetic models struggle with the tail of the distribution. Vaccines for pregnant women, orphan diseases, or uncommon polypharmacy patterns may be poorly captured. Sicne generative models rely on instance patterns, a few rare cases may be either ignored or memorized. The Synthea study noted zero instances of complications or controlled hypertension in their synthetic data (bmcmedinformdecismak.biomedcentral.com), a clear flaw. Acceptable synthetic data for PV/clinical must explicitly address such gaps: one cannot simply ignore rare but important outcomes. Possible solutions include integrating external clinical guidelines or expert rules about rare events into the model (as the authors of [5] suggest) or using hybrid methods (e.g. merging some real rare-case records into the synthetic set under strict protocols).

- **Validation Burden:** The literature repeatedly emphasizes that **validation frameworks are lacking** (jamanetwork.com) (journals.plos.org). Without standard benchmarks or regulatory guidance, each synthetic project currently has to define its own metrics and justification. This places a heavy burden on researchers and sponsors. Multiple sources call for unified standards and best practices (jamanetwork.com) (journals.plos.org). For now, the safest approach is to follow published examples rigorously: compare dozens of statistical features, replicate actual analytic results, and document every assumption.

- **Regulatory Hesitancy:** Finally, one must acknowledge cultural inertia. Regulators and ethics boards are traditionally conservative about data integrity. Synthetic data will likely face initial skepticism. The case of "fake control arms" illustrates this: as of late 2024, both FDA and EMA have *not* approved any medical application based solely on an artificially generated cohort (www.pnas.org). The UK's MHRA has publicly said they have *"not accepted a pure synthetic data control arm"* (www.pnas.org). Thus, acceptance criteria must often be more stringent than purely technical; synthetic data might only be accepted as *supplementary* evidence or in simulation spin-off analyses, until confidence grows.

Future Directions and Recommendations

Looking ahead, several developments are likely to shape acceptance criteria and practice:

- **Regulatory Guidance:** We anticipate formal guidelines clarifying synthetic data use. For example, the EU's **Artificial Intelligence Act** (enforced 2024) acknowledges synthetic data as a category in AI systems (jamanetwork.com), but has yet to set privacy standards. Guidance documents (e.g. from ICH or FDA's novel trial section) may begin to address how to integrate synthetic datasets. The EU recently initiated a privacy guideline for synthetic data by the Public Sector Information Team (PET) to provide risk-mitigation steps (jamanetwork.com). Similar frameworks could emerge in pharma. Engagement between industry and regulators (e.g. joint workshops) is already discussing how synthetic data might be handled in dossiers.

- **Benchmark Datasets and Frameworks:** The field needs standardized synthetic datasets (publicly available) to compare methods. Just as ImageNet benchmarked computer vision, we might see a synthetic healthcare data challenge. Proposals like the ATEN framework (journals.plos.org) and tools like HAPNEST (pmc.ncbi.nlm.nih.gov) point the way. We expect to see more open collections of synthetic patient data, possibly under initiatives like the European Health Data Space (EHDS) (pmc.ncbi.nlm.nih.gov), which may release synthetic samples for R&D use.

- **Integration with Privacy Tech:** Synthetic data generation will be combined with privacy-enhancing technologies. For instance, *differential privacy* during GAN training can give formal privacy guarantees (at the cost of some utility). *Federated learning* might train synthetic models without central data pooling. Combining synthetic data with techniques like homomorphic encryption or secure multi-party computation could further reduce need for raw data sharing. The JAMA article suggests that hybrid approaches (federated learning + synthetic data) are promising (jamanetwork.com).

- **Improved Metrics and Audits:** As more synthetic datasets are generated, best practices for auditing will coalesce. We expect to see consensus on certain metrics (e.g. median absolute deviation of priors, risk scores comparisons) to declare a synthetic set "fit for use" in PV/clinical. Tools that quantify privacy leakage (e.g. membership inference attacks) will be routinely applied as part of validation. "Synthetic readiness" certification (analogous to software validation) might even become a service offered by CROs or auditors.
- **Ongoing Monitoring:** An important distinction between traditional trials and synthetic uses is that synthetic models can evolve. Accepting synthetic data may require ongoing model stewardship: e.g. periodic re-validation against new data, monitoring of drift, and user feedback loops. This dynamic aspect means acceptance is not a one-time checkbox but a process.

Conclusion

Synthetic data hold enormous promise for pharmaceutical R&D, from improving drug safety surveillance to accelerating clinical insights. When properly generated and validated, synthetic health datasets can unlock new research opportunities while preserving patient privacy (pmc.ncbi.nlm.nih.gov) (www.frontiersin.org). However, the bar for "surviving inspection" is high. This report has shown that acceptance criteria must be multi-dimensional: synthetic data must statistically mirror real data (fidelity), produce the same conclusions in analysis (utility), and maintain strict confidentiality guarantees (privacy) (ascopubs.org) (jamanetwork.com). Meeting these criteria demands careful methodology and transparent validation, as exemplified by recent case studies (ascopubs.org) (pubmed.ncbi.nlm.nih.gov).

Currently, no single standard or guideline fully governs synthetic clinical data. Stakeholders must rely on best practices from the literature and lessons learned. For example, using synthetic data for a PV algorithm should involve benchmarking signal detection on known data (jamanetwork.com), while any synthetic arm for trial analysis must replicate efficacy outcomes within predetermined tolerances (ascopubs.org). Until formal regulations catch up, it is prudent to treat synthetic datasets with the same rigor as real data: thoroughly document the generation process, publish comparative metrics, and obtain ethics/regulatory review.

The outlook is optimistic: leading analysts project that in a few years *most* training data for AI in health could be synthetic (ascopubs.org). With evolving regulatory frameworks (e.g. EU AI Act) and growing community expertise (journals.plos.org) (ascopubs.org), synthetic data will likely become an accepted asset rather than a novelty. Ultimately, the true test is utility: as one perspective noted, if synthetic data can "fine-tune and reduce safety concerns even further," it may well become indispensable (www.pnas.org). By rigorously defining and applying acceptance criteria now, pharma can ensure this transition benefits patient safety and scientific integrity alike.

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies including Scilex Holding Company (SCLX) and leading CROs across North America.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.