Clinica Chimica Acta 585 (2026) 120878

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/cca

Synthetic data in the clinical laboratory: methods, applications, and future prospects

Tahir S. Pillay [a,b,*], Barbara S. van Deventer [a], Siphokazi Gwiliza [a], Evette L. Subramoney [a], Chantal van Niekerk [a]

[a] Department of Chemical Pathology, Faculty of Health Sciences, University of Pretoria and National Health Laboratory Service Tshwane Academic Division, Pretoria, South Africa
[b] Division of Chemical Pathology, University of Cape Town, South Africa

A B S T R A C T

Clinical laboratories face stringent privacy constraints, limited datasets for rare conditions, and rising demands to validate AI algorithms and workflows safely. Synthetic data—artificially generated data that preserve the statistical characteristics of real clinical data without exposing patient identities—has emerged as a powerful tool to address these challenges. This review provides a comprehensive overview of synthetic data in the context of laboratory medicine. We begin by defining synthetic data and describing the main generation methods, from rule-based simulations to modern generative models (including generative adversarial networks, variational autoencoders, and diffusion models) with examples of their use in healthcare. We then delve into key applications in the clinical laboratory: quality control and method validation, education and training, machine learning development, test utilization and workflow simulation, and external quality assessment. Advantages of synthetic data—such as enhanced privacy, scalability, flexibility in simulating rare events, and cost-effectiveness—are discussed with illustrative case studies. We also examine challenges and limitations, including concerns about data fidelity, bias amplification, risks of model overfitting or re-identification attacks, and the cautious stance of regulators that still require real patient data for approvals. Finally, we outline future directions for synthetic data in laboratory medicine, from hybrid real–synthetic datasets and privacy-enhancing techniques to evolving regulatory frameworks and the potential to democratize data access globally. While synthetic data cannot entirely replace real clinical data—especially for regulatory validation—it can significantly augment what laboratories can design, test, and achieve, provided it is used with careful validation and ethical safeguards.

1. Introduction

Modern clinical laboratories operate under tight data-sharing restrictions and often contend with small or skewed datasets. Patient privacy regulations like the Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) limit the exchange of real patient data across institutions, and rare diseases or unusual test scenarios may be under-represented in data from a single laboratory [1]. At the same time, there is an increasing need to develop and validate complex algorithms (from middleware and laboratory information systems to artificial intellence (AI)-driven diagnostic tools) before deploying them in high-stakes patient care [2]. Together, these factors create a demand for alternative data sources that are realistic yet pose no privacy risk to patients. Synthetic data offers a compelling solution: it allows laboratories to generate unlimited realistic records or results "in silico" to supplement real data, test hypotheses, and train models without exposing any actual patient information [3].

Interest in synthetic clinical data has surged over the past decade. Early demonstrations showed that sharing synthetic electronic health records (EHRs) could mitigate privacy concerns while still enabling research and machine learning (ML) development [4]. Recent reviews highlight a proliferation of techniques to create high-fidelity synthetic health data across modalities (tabular records, time-series signals, medical images, etc.) and an expanding list of use cases in healthcare [2]. Notably, a working definition from the Alan Turing Institute and the Royal Society describes synthetic data as *"data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a data science task"* [2]. In practice, synthetic data generation in medicine ranges from simple simulations to advanced generative AI; however, the field lacks a single consensus definition, leading to inconsistent use of the term in different contexts [2]. This ambiguity is gradually being addressed as researchers propose more precise taxonomies and standards for describing synthetic datasets [5].

* Corresponding author at: Department of Chemical Pathology, Faculty of Health Sciences, University of Pretoria and National Health Laboratory Service Tshwane Academic Division, Pretoria, South Africa.
E-mail address: tahir.pillay@up.ac.za (T.S. Pillay).

https://doi.org/10.1016/j.cca.2026.120878
Received 5 January 2026; Received in revised form 29 January 2026; Accepted 29 January 2026
Available online 30 January 2026

Early adopters in finance and technology demonstrated that synthetic data can unlock data sharing and algorithm testing when real data are scarce or sensitive [2]. In laboratory medicine, the potential is similarly large but accompanied by caution. Stakeholders (laboratory directors, clinicians, regulators) are understandably risk-averse; they demand evidence that synthetic data faithfully represents reality and that models trained on such data generalize to real patients [2]. There is also the paramount concern of patient safety: any insights or tools derived from synthetic data must be rigorously validated against real-world outcomes before clinical implementation [2]. With these precautions in mind, we examine how synthetic data is generated and used in the clinical laboratory today including what benefits it offers, what pitfalls exist, and how future developments might further integrate synthetic data into routine laboratory science and quality management.

2. Definition and generation of synthetic data

Synthetic data refers to artificial datasets that are generated to resemble real-world data in statistical properties and structure, without directly using identifiable real records [1]. In other words, synthetic data aims to mimic the distributions, correlations, and patterns found in genuine clinical data (e.g. laboratory test values, patient demographics, instrument readings) while containing *no actual patient information*. According to the U.S. Food and Drug Administration (FDA) digital health glossary, synthetic data consists of "new values and/or data elements that are generated *artificially* (e.g., through statistical modeling, computer simulation)" such that they are statistically similar to real data but not tied to any real individual [5]. This is distinct from de-identified data, which are directly derived from actual patients and therefore carry a (hopefully small) risk of re-identification, whereas, fully synthetic data, by contrast, are *built from scratch* and, if properly generated, cannot be traced back to any one person [3].

2.1. Types of Synthetic Data and Generation Methods (Table 1)

There are multiple approaches to generate synthetic data, falling into three broad families [1]: Rule-based simulations, Statistical and probabilistic modelling and machine learning-based generative models.

2.1.1. Rule-based simulations

These use explicit domain knowledge and predefined rules or equations to create data. For example, a mathematical model of glucose and insulin dynamics could be used to produce synthetic blood glucose readings over time, given virtual patient parameters. In healthcare, an emblematic tool is *Synthea*, an open-source patient simulator that models entire synthetic patient records by simulating diseases, treatments, and outcomes [6]. Synthea uses disease progression rules and module-based logic to generate realistic sequences of events (e.g. clinic visits, lab tests, hospitalizations) for fictional patients. One million such synthetic patient records (complete with demographics, laboratory test results, diagnoses, and clinical notes) have been made freely available via Synthea in standard health data formats [6]. Rule-based synthetic data is essentially a "virtual clinical trial" or "digital twin" approach – creating simulated patients or laboratory scenarios according to expert-defined parameters. It can be extremely useful for modeling specific *what-if scenarios* (for instance, how an outbreak might spread and be reflected in laboratory results) but may be limited by the completeness and accuracy of the rules encoded.

Table 1
Summary of methods for the generation of synthetic data.

Method	Strengths	Limitations	Example tools
Rule-based Simulations	Transparent, domain knowledge-driven; good for 'what-if' scenarios	Limited by completeness of rules; may oversimplify reality	Synthea
Statistical/ Probabilistic Models	Captures correlations; flexible for tabular data	Struggles with high-dimensional data; requires careful tuning	Bayesian Networks, Copulas
GANs	High realism, adaptable to images/ tabular; useful for rare cases	Mode collapse; requires large datasets and compute	medGAN, CTGAN, DCGAN
VAEs	Stable training, interpretable latent space, uncertainty estimates	Blurry outputs; less sharp detail than GANs	VAE-GAN hybrids

2.1.2. Statistical and probabilistic modeling

These methods rely on *learned probability distributions* or statistical relationships inferred from real data to generate new samples. Traditional techniques include extracting values from fitted probability distributions, *bootstrapping* (resampling with replacement from an empirical dataset), or using probabilistic graphical models. For example, Bayesian networks have been used to synthesize health records by capturing conditional dependencies between variables (e.g., the relationship between age, laboratory results, and diagnosis) [1]. In one study, a Bayesian network approach was employed to generate high-fidelity synthetic primary care patient data in the United Kingdom, yielding data that preserved key multivariate correlations from the original dataset [7]. Another example is the use of copulas or multivariate distributions to produce synthetic laboratory result datasets that maintain realistic covariance structures. These statistical approaches are generally more *flexible* than simple rule-based simulations since they can ingest real data and attempt to reproduce its patterns. However, they often require careful tuning and may struggle with very complex, high-dimensional data unless combined with more advanced techniques.

2.1.3. Machine learning-based generative models

In recent years, data-driven AI models have become the state-of-the-art for generating synthetic data [1,8]. These include deep learning approaches such as Generative Adversarial Networks (GANs) [1,9], Variational Autoencoders (VAEs) [1] and more recently diffusion models [10]. GANs, first introduced in 2014, consist of a two-network system (generator and discriminator) trained in opposition: the generator tries to create fake data that the discriminator cannot distinguish from real data. Through this adversarial process, GANs can learn to generate highly realistic outputs. Many GAN architectures have been developed for different data types: for example, *medGAN* was an early GAN tailored to multi-label discrete clinical data, capable of generating synthetic patient records of diagnoses and procedures [4]. Choi et al. demonstrated that medGAN's synthetic records maintained similar statistical properties to real EHR data and yielded machine learning models with predictive performance comparable to models trained on real data [4]. They also reported only limited privacy risks (low likelihood of re-identifying actual patients) in medGAN's outputs [4]. Building on such work, more specialized GANs have emerged: Conditional Tabular GAN (CTGAN), for instance, was designed to handle mixed data types in tables and to faithfully model rare categorical values by sampling conditional distributions [1]. In one case, CTGAN was used to generate synthetic clinical datasets for decision support research, and the resulting data passed various quality tests (comparison of distributions, pairwise correlations) against the real data [7]. For images, *deep convolutional GANs (DCGANs)* can produce high-resolution synthetic medical images such as pathology slides [1] whereas *conditional GANs (cGANs)* allow control over outputs (e.g. generate pathology images *with* a specified tumor type) [1]. A recent pathology study used a cGAN to create "deepfake" histology images of prostate and colon tissues conditioned on cancer type; the synthetic images accurately reproduced morphological features of both common and rare tumor subtypes [11].

Notably, board-certified pathologists were unable to distinguish many of these synthetic histology images from real specimens, and when the synthetic images were used to train a diagnostic AI system, the performance was as good as training on real data [11]. Such results showcase how far generative models have advanced in creating realistic synthetic data.

Beyond GANs, VAEs provide an alternative deep generative approach. VAEs learn a compressed latent representation of real data and then sample from this latent space to generate new outputs. These tend to be more stable to train and can naturally incorporate uncertainty estimates. However, their images may be blurrier than GANs [1]. Nonetheless, VAEs and GANs can be combined (in a VAE-GAN hybrid) to leverage the strengths of both [1]. Meanwhile, diffusion models have recently gained attention for synthetic data generation. Diffusion models work by iteratively perturbing data with noise and then learning to reverse that process, essentially "denoising" to generate new data samples. Originally very successful in image generation, diffusion approaches have been adapted to tabular data (e.g., *TabDDPM*) and have shown superior fidelity to GAN/VAEs on several benchmarks [12]. Kotelnikov et al. introduced TabDDPM, a diffusion model for mixed-type tabular data, and demonstrated that it outperformed GAN-based and VAE-based methods in capturing complex feature relationships [12]. Importantly, they noted that TabDDPM's synthetic data were preferable to simpler oversampling (SMOTE) for *privacy-sensitive scenarios*—since no real record is replicated, synthetic data could be shared publicly without breaching confidentiality as with real data [12]. This highlights how advanced generative models can produce high-utility data *and* offer privacy protection.

It should be noted that synthetic data methods are not mutually exclusive. Often a hybrid approach is used: for example, a statistical model might generate base data which is then refined by a GAN, or real data might be augmented with synthetic samples (as in "semi-synthetic" or partially synthetic datasets). Indeed, synthetic data exists on a spectrum from *fully synthetic* (all records are artificially generated with no one-to-one correspondence to real cases) to *partially synthetic* (some real data is retained or some fields are synthesized) [1,2]. There is also the concept of hybrid synthetic data, where an entire synthetic dataset is blended with some proportion of real data records [5]. According to one taxonomy, partially synthetic data (replacing only selected sensitive variables in an otherwise real dataset) can still carry re-identification risk, whereas fully synthetic data eliminates direct identifiers but may sacrifice some utility; a hybrid approach (mixing real and synthetic records) can offer a balance of strong privacy protection while maintaining higher fidelity on key variables [5]. The choice of method depends on the use-case and risk tolerance.

2.2. Data modalities and practical method–data fit

Clinical laboratory data are heterogeneous, and the most appropriate synthesis approach depends strongly on the data modality and the downstream task. Below we link common laboratory data modalities to generation approaches typically best suited to each, with practical caveats for readers new to the field.

(I). Tabular laboratory datasets (e.g., chemistry/immunoassay panels): These are well suited to statistical/probabilistic methods (e.g., copulas, Bayesian networks) and tabular deep generative models (e.g., CTGAN/TVAE-style architectures) [13,14]. Key aims are preserving inter-analyte correlations, constraints (physiologic ranges, reportable limits), and clinically important tail behavior.

(II). Longitudinal/time-series results (e.g., Patient-based Real Time Quality Control Streams (PBRTQC) streams): These can be synthesized using time-series extensions of statistical models (e.g., state-space/ Vector Autoregression (VAR)-type approaches), sequence autoencoders, or time-series GAN variants. Rule-based 'scenario injection' is often highly useful here to create controlled step shifts, gradual drift, seasonality, and outliers for stress-testing detection algorithms [15].

(III). Images (e.g., digital pathology): Images typically require high-capacity deep generative models (GANs or diffusion models) for realistic image synthesis/augmentation [16]. Careful validation is essential to ensure clinically relevant morphology is preserved and artifacts are not introduced; rule-based methods are generally unsuitable for high-dimensional image generation.

(IV). Text (e.g., Laboratory information system (LIS) free-text/ reporting elements):This may be generated using template-based approaches for constrained outputs or transformer/(Large Language model (LLM)-based models for richer narratives [17]. Since text models can inadvertently reproduce memorized fragments, governance should include stringent de-identification, disclosure risk assessment, and human review—especially if datasets were derived from identifiable sources [18].

In practice, 'method–data fit' is also shaped by the intended use (eg. education/training, algorithm development, PBRTQC stress testing, or operational simulation). For many laboratory applications, a hybrid strategy (e.g., rules/constraints layered onto statistical or ML generators) provides the best balance of realism, interpretability, and risk control.

In summary, synthetic data can be generated by a variety of techniques ranging from simple simulations to cutting-edge AI. Generators like Synthea exemplify rule-based simulations producing complete mock health records [5], while models like medGAN and CTGAN demonstrate data-driven approaches for discrete health data [1] [4] and diffusion models like TabDDPM push the frontier for high-dimensional tabular synthesis [12]. Regardless of method, a critical aspect is evaluation: one must verify that synthetic data are statistically similar to real data (for utility) *and* that they do not inadvertently leak information about any real individuals (for privacy). Various quality metrics have been proposed, such as distributional similarity tests, machine learning efficacy (can models trained on synthetic data perform well on real test sets?), and privacy metrics like membership inference attack resistance [1]. We will discuss these concerns later in the context of challenges. First, we will examine how synthetic data is being applied in the clinical laboratory setting.

3. Applications in the clinical laboratory

Synthetic data has diverse applications across the clinical laboratory domain, touching many aspects of operations and research. Here we expand on several key use cases (Table 2):

3.1. Quality control and method validation

One foundational use of synthetic data in laboratories is to support analytical quality control (QC) and the validation of new methods or instruments. In traditional QC, laboratories rely on limited pools of control materials or archived patient specimens to assess whether an assay is performing within tolerance. Synthetic data offers the ability to *simulate large volumes of test results* under varying conditions, thereby stress-testing QC procedures without needing endless real samples. For example, a LIS developer can generate millions of synthetic patient results with known error patterns to evaluate whether the LIS delta-check algorithms or autoverification rules correctly flag abnormalities [19]. Such broad testing is often impractical with real data but easy to do with synthetic results that cover every edge-case scenario (extreme high/low values, rare but critical result combinations, etc.).

Synthetic data has a particularly valuable role in PBRTQC [20–23]. PBRTQC relies on continuous monitoring of routine patient results to detect analytical shifts, but its effectiveness can be limited when rare pathological cases or small sample sizes distort the baseline distribution.

Table 2
Clinical Laboratory applications of synthetic data (benefits, challenges, and implementation considerations).

Domain	Typical uses	Benefits	Challenges/risks	Regulatory / implementation
QC / PBRTQC	Simulate patient-result streams; inject step shifts, drift, outliers; benchmark PBRTQC; tune alert thresholds; staff training.	Controlled 'what-if' experiments; improves robustness and comparability of detection methods; reduces need to share patient-level data.	May miss rare failure modes; generator-driven bias; overfitting PBRTQC to synthetic patterns; needs periodic re-validation as drift occurs.	Define intended use (development vs clinical); version/ change-control; validate against real data; disclosure risk testing if derived from identifiable cohorts.
Education / training	Case sets; competency assessments; simulated LIS datasets; onboarding; exam preparation.	Scalable and shareable learning datasets; safe exposure to realistic distributions and edge cases; supports cross-site training.	Pedagogic bias if generator reflects local practice; unrealistic cases can mislead; requires curation and clinical plausibility checks.	Governance approval; label clearly as synthetic; align with curriculum outcomes; maintain audit trail and updates.
AI development	Augment rare conditions; prototype/ benchmark models; facilitate collaboration where data-sharing is constrained.	May improve generalization when real data are scarce; accelerates iteration; supports multi-site collaboration.	Utility depends on fidelity; synthetic bias can amplify disparities; leakage/memorization risk; domain shift at deployment.	Use synthetic primarily for training/experimentation; require external validation on real cohorts; track provenance/ versioning; monitor fairness and drift post-deployment.
Test utilization	Simulate ordering patterns; evaluate stewardship interventions; demand forecasting; anomaly detection.	Safe sandbox for scenario testing; supports capacity planning and utilization stewardship without impacting care.	Ordering behavior is context-dependent; may omit drivers (policies, incentives); confounding may be misrepresented.	Treat as decision-support development; align with institutional governance; evaluate interventions prospectively with appropriate ethical oversight.
Workflow simulation	End-to-end lab simulation (arrival→TAT); staffing/instrument capacity; downtime contingency planning.	Identifies bottlenecks; supports service-level planning and resilience; improves operational decision-making.	Sensitive to assumptions and parameter misspecification; may under-represent human factors and local constraints.	Integrate with Quality Management system (QMS); document assumptions; validate against historical operational metrics; change-control for updates.

Synthetic patient profiles can be generated to simulate thousands of realistic test results, including controlled biases or drifts, allowing laboratories to benchmark different PBRTQC algorithms (Fig. 1, [20–23]), evaluate optimal block sizes, and test sensitivity for bias detection without risking patient safety. By enriching the dataset with rare or extreme cases, synthetic data strengthens the robustness of PBRTQC design and validation, ensuring that QC systems remain both sensitive and specific under diverse clinical conditions.

Synthetic datasets are also valuable for method validation studies, particularly when a new assay is introduced but only limited patient data are initially available. Researchers can create a synthetic cohort that mirrors the demographics and basic laboratory values of a small real dataset, then use this expanded cohort to statistically derive or test performance thresholds. A concrete example comes from the validation of a novel high-sensitivity troponin point-of-care (POC) assay for emergency cardiac screening. Pickering et al. (2024) combined real patient data with synthetically generated patient results to determine optimal cut-off values for ruling out acute myocardial infarction (AMI) [24]. In their study, only ~190 patients had confirmed AMI, which made it hard to set a reliable low-level troponin threshold. By generating 500 synthetic datasets based on the distribution of troponin results and covariates, they were able to repeatedly simulate the scenario and identify a cutoff (5 ng/L) that would achieve 99% sensitivity for rule-out [24]. This approach essentially *augmented* the sample size via synthesis, leading to more robust estimates than the real data alone could provide. The authors concluded that synthetic data generation was an effective way to derive safe screening thresholds and demonstrated the feasibility of integrating synthetic data into assay validation workflows [24].

Similarly, synthetic data can aid instrument manufacturers in lot verification and calibration. Before rolling out a new lot of reagents or a software update in an analyzer, a manufacturer could simulate thousands of test runs under specified conditions to ensure the QC flags behave as expected. This kind of *in silico* validation enables the identification of potential issues early. During the COVID-19 pandemic [25], for instance, synthetic patient data models played a role in evaluating testing protocols when real samples were limited. Walonoski et al. developed a Synthea-based COVID-19 disease progression model that generated synthetic patients and laboratory results for scenario analysis [7]. Synthetic COVID-19 data enabled investigators to test the impact of various screening strategies and to train predictive models at a time when datasets from actual patients were sparse in the pandemic's early phase.

In summary, synthetic data extends quality assurance capabilities by offering unlimited test scenarios. Laboratories can simulate high-dimensional combinations of results (including rare outliers) to evaluate the robustness of their QC rules, reference ranges, and alert systems are robust. This reduces the dependence on waiting for rare real-world errors or pathology to occur. Moreover, because synthetic QC data contains no real patient information, it can be freely shared among laboratories or with vendors to collaboratively improve systems without privacy hurdles. As one simulation study noted, using synthetic patient data to evaluate point-of-care cardiac marker protocols provided insights that would have taken *years* to obtain from clinical observation, and the synthetic approach allowed complete reproducibility and data sharing "without the need for ethical approval" [26]. This illustrates how synthetic data can accelerate method development while upholding compliance and patient safety.

3.2. Training and education

Clinical laboratories are also leveraging synthetic data to enhance education and staff training. In teaching environments, synthetic case data can serve as realistic practical material for laboratory scientists, pathology residents, and other trainees [27,28]. Because the data are not real, there is no risk of violating patient privacy when incorporating them into classroom exercises, examinations, or competency

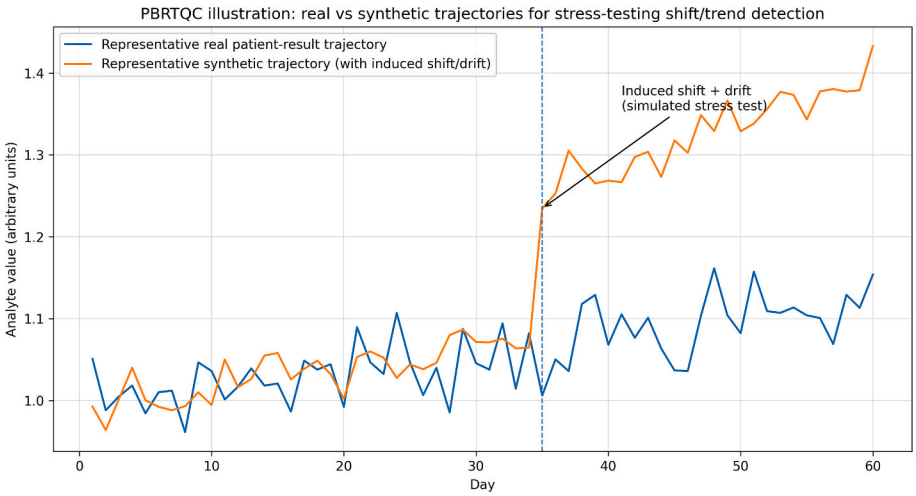

Fig. 1. Application of a synthetic data approach to PBRTQC.

assessments. For example, an instructor can generate a synthetic dataset containing a mix of normal and abnormal laboratory results, including some results indicative of rare diseases that a trainee might never encounter during a short rotation. The trainees can then practice interpreting the data or troubleshooting QC failures in a simulated setting. Synthetic case vignettes have been used to teach interpretation of QC rules (like Westgard rules for detecting assay drift) by providing numerous scenarios of control source data runs, some embedded with shifts or trends, for students to identify. This is far more efficient than waiting for such QC events to occur in real time.

Synthetic data is particularly valuable for experiential learning of rare events. A training module on inborn errors of metabolism, for instance, could include synthetic newborn screening results that mimic those of an infant with an extremely uncommon metabolic disorder. Trainees can then learn to recognize the abnormal pattern (e.g., specific enzyme assay values) without needing an actual case on hand. In pathology education, generative models now enable the creation of synthetic pathology images ("digital slides") showing rare malignancies or subtle morphologic variants. Falahkheirkhah et al. (2023) demonstrated a GAN-based system that produces *"deepfake" histology images* for digital pathology training [11]. Their framework allowed users to generate new histologic variations on demand – effectively creating unlimited teaching slides for even the rarest morphologies. This kind of tool can make pathology residents more familiar with variability in tissue appearance and ensure exposure to edge cases that may not be typically present in a hospital archive [11].

Professional organizations have begun to recognize the importance of synthetic data in education. The College of American Pathologists (CAP) now includes sessions on synthetic data generation and its applications in pathology informatics curricula [29]. In these sessions, trainees not only learn how synthetic data are created but also how they can be used to bridge gaps between what AI algorithms can do and what real-world data is available. Synthetic data can also help with competency assessments: laboratories can test assess staff interpretive skills using fictitious yet realistic case data, which can be generated fresh to prevent memorization. One could envision a proficiency test for laboratory technologists where all the patient results are synthetic – ensuring that all examinees get an equivalent challenge, and no actual patient results are disclosed.

Another angle is using synthetic data to train data scientists and informaticians in the laboratory. Working with real health data often requires secured access, but synthetic datasets (e.g., a synthetic EHR database complete with laboratory results and outcomes) can be open-access [7,30]. This enables data science students to develop and test clinical laboratory analytics without needing IRB approval or business associate agreements. A recent review noted that by leveraging synthetic data, researchers can access "high-quality, representative multimodal datasets without exposing sensitive patient information," thereby accelerating training of AI models and *education of practitioners* alike [7]. In pathology and laboratory medicine, Pantanowitz et al. (2024) similarly observed that synthetic data can "enhance our medical education and research/quality study needs" by providing abundant, diverse examples for learning and quality improvement drills [31].

In summary, synthetic data empowers education by providing safe, shareable, and customizable teaching materials. Trainees can gain experience with scenarios that would be rare or impossible to assemble from real cases, from unusual laboratory results to exotic pathology images. This not only improves competency but does so in a manner that respects patient privacy and data regulations. As generative technology advances, we can expect digital simulators and virtual datasets to play an increasingly prominent role in laboratory medicine training.

3.3. AI and machine learning model development

Perhaps the most active area for synthetic data in the clinical laboratory is in the development of ML and AI models. High-quality ML models typically require large and diverse datasets, which in healthcare can be a bottleneck due to limited sample sizes and privacy restrictions. Synthetic data can alleviate these issues in several ways:

3.3.1. Data augmentation and balancing

Laboratory datasets often suffer from class imbalances or underrepresentation of certain demographics. For example, an algorithm to detect atypical cells in blood smears might have thousands of examples of normal cells but very few images of a rare leukemia. Generative models can be used to create additional synthetic examples of the minority class (e.g., synthetic images of the rare leukemia cells) to balance the training data [7]. By augmenting the real dataset with these generated cases, the model can learn more generalized features and avoid being biased toward the majority class. In one study aiming to improve ML predictions for under-represented patient groups, researchers employed a toolkit of generative techniques (including *RadialGAN*, *TabDDPM*, and *CTGAN*) to oversample minority populations in a clinical dataset [7]. This approach, made available via an open-source library (*SynthCity*), improved model accuracy for those minority groups, illustrating how synthetic data can directly tackle data skew and fairness issues.

3.3.2. Improving model robustness

Models trained on limited real data may overfit or fail to generalize

to new settings. By training or pre-training models on a wide range of synthetic data variations, one can potentially make them more robust. For instance, a digital pathology AI could be exposed to a broad synthetic image dataset that covers variations in staining intensity, imaging artifacts, and tissue aberrations, such that the model becomes invariant to those factors when it later sees real images. Generative synthetic data can introduce controlled noise and perturbations to challenge the model during training, acting as a form of *simulation-based stress testing* for the algorithm.

3.3.3. Facilitating data sharing and multi-institutional studies

Perhaps one of the greatest benefits of synthetic data for AI development is the ability to share and combine data across institutions without violating privacy. Under HIPAA and GDPR, sharing actual patient-level data between laboratories or companies is arduous and often impossible unless the data are de-identified to strict standards. Fully synthetic datasets, however, contain no real patient identifiers or records, so in principle they can be shared freely as they are not considered protected health information [1]. This opens the door for collaborative development of ML models: for example, multiple hospitals could contribute to a common synthetic dataset reflecting a rare disease, enabling an AI model to be trained on a dataset larger than any single site's real-world data. Researchers have noted that synthetic health data enables cross-institution collaboration "while ensuring compliance with stringent privacy laws" [1]. As a case, the National Institutes of Health's National COVID Cohort Collaborative (N3C) created a large repository of COVID-19 clinical data, and to broaden access they also released a parallel *synthetic* dataset generated from the real data. This allowed many researchers to experiment and build preliminary models without touching the actual patient records, speeding up innovation while maintaining privacy safeguards [1].

3.3.4. Overcoming data scarcity for rare conditions

In diseases or laboratory scenarios where collecting a substantial real dataset is extraordinarily difficult, synthetic data may be used to *bootstrap* model development. Rare genetic disorders, unusual laboratory test interference cases, or outbreak scenarios (like an emerging pathogen) fall into this category. As a perspective article on rare diseases pointed out, synthetic data can provide "diverse and privacy-preserving datasets" that enable AI models to be trained for detecting rare markers or conditions, thereby accelerating innovation in diagnostics where real data are too scarce [1]. For example, an AI model for newborn screening of a metabolic disease that only has a few dozen real cases worldwide might begin by training on thousands of synthetic cases generated by a biochemical simulation of the disease condition laboratory profile, ensuring the model is exposed to a wide spectrum of possible presentations. This synthetic pre-training can later be fine-tuned with whatever real cases exist.

3.3.5. Testing algorithms under virtual trials

Beyond training models, synthetic data allows AI algorithms to be evaluated *in silico* before prospective trials. One can create a synthetic patient population (complete with laboratory values and outcomes) to serve as a testbed for an AI-driven decision support system. The AI's recommendations can then be evaluated against the known synthetic "ground truth" outcomes to estimate how it might perform in real practice. This kind of internal validation with synthetic cohorts can flag problems in the algorithm or areas where it extrapolates incorrectly, all before any real patient is involved.

Numerous studies underscore the above-mentioned benefits. The medGAN work described by Choi et al. [4] was seminal in showing that models trained on synthetic patient record data achieved performance on par with those trained on real data for tasks like risk prediction [4]. More recent investigations in digital pathology have shown that supplementing real training images with *GAN-generated synthetic images* improved diagnostic model accuracy and generalization, especially for rarer conditions [11]. In time-series analytics (e.g., physiological signals, laboratory result trends), models like TimeGAN have been used to generate plausible patient trajectories that aid algorithms learn temporal patterns without accessing real longitudinal patient data [1]. Moreover, using synthetic data can mitigate the need for data labeling by humans; if the synthetic generator can also output labels (which it can, since it knows the class of each generated sample by design), one is able to create massive labeled datasets at will. This is particularly useful for training supervised ML models in the laboratory (for instance, an image classifier to detect malarial parasites can be trained on synthetic blood smear images that are automatically labeled by the generation process, sidestepping the need for a pathologist to label thousands of training images).

Another practical outcome of using synthetic data is model transparency and debugging. Because one has complete knowledge of the synthetic data generation process, one can construct specific scenarios to probe the behavior of the model. For example, one could systematically vary a single laboratory value in a synthetic patient profile across a range and see how the AI's risk prediction changes, thereby illuminating the model's reliance on that value [32]. This kind of testing is easier with synthetic data than real data, where you cannot arbitrarily set or change values without losing realism or violating ethical use of patient information.

In summary, synthetic data is accelerating the development of ML models in laboratory medicine by providing *volume, variety,* and *shareability* of data that would otherwise be unattainable. It enables oversampling of under-represented cases to reduce bias [7] supports safe multi-center data pooling in compliance with privacy regulations [1] and offers a sandbox for algorithm development and validation without patient risk. The result is faster and more inclusive progress in applying AI to laboratory problems, from automated image analysis in pathology to predictive analytics in clinical chemistry. In the next section, we consider how synthetic data aids the design of laboratory processes and external evaluations.

3.4. Test utilization and workflow simulation

Clinical laboratories continually seek to optimize how tests are used (utilization) and to anticipate the effects of changes in testing strategies on operations. Synthetic data can be an invaluable tool for simulation of laboratory workflows and test utilization scenarios. Instead of implementing a change in the real world and observing over months how it unfolds, laboratories can simulate it with synthetic data in a matter of hours.

One application is in modeling the impact of introducing a new test or changing testing protocols. For example, when high-sensitivity cardiac troponin assays were first introduced, emergency departments and laboratories had to adjust protocols for ruling out AMI. By creating a synthetic cohort of patients presenting with chest pain – with troponin results generated according to known distributions for both confirmed AMI and non-AMI patients – one can simulate how different rule-out algorithms perform. Robinson et al. (2009) conducted such a simulation study using synthetic patient data to compare various accelerated diagnostic protocols for chest pain in the emergency setting [26]. They generated synthetic patient records that matched the statistical characteristics of actual patients from the emergency department(ED) patients (drawing on a seed dataset of 137 cases) and then virtually applied multiple POC troponin testing protocols to these synthetic patients [26]. The simulation revealed the sensitivity trade-offs at different time cut-offs (e.g., 90-min vs 120-min troponin recheck) and demonstrated that some protocols could achieve over 98% sensitivity for rule-out within 2 h [26] Crucially, the authors noted that this level of analysis would have been impractical to do with real patients alone, as it would require years of data collection to get comparable numbers of cases under each scenario [26]. Synthetic data allowed rapid "A/B testing" of clinical pathways and provided evidence that could inform protocol selection [26].

Laboratories can similarly model test ordering patterns and downstream effects. Suppose a hospital is considering implementing action of reflex testing (automatic add-on tests based on initial results) to reduce unnecessary follow-ups. A synthetic utilization dataset can be generated by simulating physician ordering behavior and patient populations, and then running the reflex rules to see how often they trigger and how much additional testing or cost is incurred. By adjusting parameters (e.g., prevalence of certain conditions, thresholds for reflex), analysts can optimize the strategy before actual rollout. This approach can also identify unintended consequences, like too many false-positive reflex triggers, which can then be mitigated by tweaking the criteria in silico.

Another area is workflow simulation and capacity planning. Laboratories use sophisticated automation and staff scheduling to handle test volumes. Synthetic time-stamped data of test orders and results can be used to simulate laboratory operational flow under different conditions. For instance, synthetic daily logs can be generated to represent a surge in testing (perhaps due to a flu outbreak or a public health screening campaign) and input into a discrete-event simulation of the laboratory automation line. This helps to identify where bottlenecks would occur and promote consideration of whether additional instruments or staffing would be needed to maintain turnaround time. It can also guide decisions such as determining the optimal batch size is for a certain analyzer or how a new middleware algorithm might reduce TAT by immediate routing of critical results.

Synthetic utilization data has also been proposed as a way to assess the potential outcomes of policy changes. For example, if a new clinical guideline suggests more frequent monitoring of a laboratory marker in certain patients, a hospital can simulate the effect on laboratory volume and patient outcomes. The synthetic dataset can incorporate patient profiles and simulate how following the guideline would generate laboratory orders over time, and whether that may, for example, lead to earlier detection of abnormalities (modeled via synthetic outcomes) or just additional costs.

In a broader sense, this ties into creating a "digital twin" of the laboratory – a virtual replica that can be experimented on. By calibrating the synthetic model with real baseline data, laboratories can evaluate modifications virtually. During the COVID-19 pandemic, many laboratories dramatically shifted testing menus and had to rearrange workflows; those with simulation tools could better anticipate where to allocate resources [25]. Synthetic data helps power these simulations in absence of historical precedent [25].

A specific benefit of synthetic workflow data is in rare or extreme scenario planning. Laboratories need contingency plans for surges (e.g., mass casualty events or emerging infections) that they have never encountered before. With synthetic data, one can simulate a thousand "outbreak days" or "disaster scenarios" to assess how the LIS and staff would cope – something one cannot do with real data given the rarity of disasters. It enables identification of potential failure points in the workflow (for example, the sample check-in process becomes a choke point, or certain QC processes take too long under heavy load) and addresses them proactively.

In short, synthetic data allows laboratories to model the future. It converts theoretical questions – *"What if our test volume doubles next winter?" "What if we implement algorithm X?"* – into quantitative simulations that can inform decision-making. As noted in one simulation study, synthetic laboratory data permitted "free exchange of information" and reproducible experiments on protocol efficacy that would not be feasible or ethical to do in vivo [26]. This capability is invaluable for evidence-based laboratory management, enabling laboratories to optimize utilization and workflows with less trial-and-error in live settings.

3.5. Benchmarking and external quality assessment (EQA)

External Quality Assessment (EQA), also known as proficiency testing, is a process where laboratories periodically test unknown samples provided by an external agency and compare results to ensure accuracy. Traditionally, EQA relies on real patient-derived materials or commutable samples sent to laboratories. However, there is growing interest in using synthetic *data* for benchmarking lab performance.

A proficiency testing program could conceivably provide laboratories with synthetic *datasets* (rather than vials of sample) and ask them to run their data analysis pipelines or interpretative algorithms on that data. For example, a digital pathology EQA could distribute a set of synthetic whole-slide images and require participating laboratories to diagnose them or run image analysis algorithms. Since the images are synthetic, an unlimited number of challenges can be created and there is no patient privacy issue. Similarly, a genomics proficiency test may use simulated DNA sequences [33] to test bioinformatics pipelines. This approach is still emergent, but it holds promise: it could lower costs (no need to ship physical samples globally) and make EQA more accessible, especially to laboratories in resource-limited settings that may be unable to import physical samples easily. Synthetic data can be transmitted instantaneously, and laboratories anywhere with an internet connection could participate in proficiency exercises, democratizing quality improvement efforts.

Another use of synthetic data in benchmarking is in evaluating laboratory algorithms or decision-support tools. Suppose multiple labs have developed algorithms for flagging implausible lab results (e.g., delta check systems). An external body could generate a standardized synthetic dataset containing a variety of longitudinal patient results, some with simulated errors or changes, and then evaluate each laboratory algorithm (e.g., detection rate, false positives) on that common dataset. This provides an objective benchmark and fosters competition and commitment to quality improvement. Researchers have advocated for frameworks to systematically evaluate the "*utility and fidelity*" of synthetic data generators for healthcare [1]– these same frameworks can be used to benchmark algorithms by using synthetic data as test input with known expected outputs.

Moreover, synthetic data can assist laboratories meet regulatory or accreditation requirements for validation when real sample availability is limited. For instance, if a laboratory must validate an assay across the entire reportable range but cannot practically get actual specimens at the extreme high end, it may use synthetic data or simulated samples to extrapolate performance at those high concentrations. Regulatory bodies still require real data for final proof (discussed later), but synthetic approaches can fill in gaps and strengthen the case.

In summary, synthetic data and samples are gradually being incorporated into the quality assurance ecosystem. EQA providers acknowledge that "synthetic material covers a wide range" of possibilities and is sometimes necessary to achieve thorough assessment [34]. Whether it is a vial of artificial proficiency testing serum or a digital file of synthetic patient results, the goal is the same: to determine whether laboratories ability to produce accurate results and interpretations. As technology and confidence in synthetic data grow, we may see routine proficiency challenges that are entirely synthetic, enabling more frequent and flexible benchmarking. This could especially benefit smaller and remote laboratories, who could receive proficiency test datasets electronically and improve their quality without the logistics of shipping physical specimens internationally [1]. It is an area poised for development, with pilot programs likely to emerge that leverage synthetic data to complement traditional EQA.

Having explored how synthetic data is used in practice – from improving instrument Quality Assurance (QA) to training the next generation of laboratorians – we now turn to a balanced look at the advantages this approach offers, as well as the challenges and limitations that must be navigated.

4. Advantages of synthetic data in lab medicine

The growing interest in synthetic data is driven by several clear advantages, especially relevant to clinical labs:

4.1. Privacy protection and regulatory compliance

Synthetic data can substantially reduce privacy risk compared with sharing raw patient-level data, but it is not automatically risk-free or exempt from outside healthcare privacy regulation. The residual risk depends on the synthesis method (e.g., Differentially private(DP) vs non-DP), overfitting/memorization, cohort size/rarity, available auxiliary data, and results of formal disclosure risk assessments (e.g., membership inference and linkage risk testing [35–37]. By removing identifiers and any direct link to real individuals, synthetic data can enable analyses and collaborations that would otherwise be legally or ethically off-limits [1]. For example, a group of hospitals in different countries could pool synthetic patient data to develop a unified QC algorithm, something they could not do with actual patient results due to data protection laws. The European Union's Data Governance Act (2022) explicitly cites synthetic data generation as a "privacy-preserving method" to facilitate data sharing in a more privacy-friendly way [5] Importantly, because synthetic data are generated artificially, they are not subject to the same legal status as medical records – if done properly, they may be considered *non-human-subject data*. Some ethicists have argued that sufficiently non-identifiable synthetic data should fall outside the scope of HIPAA altogether [1,34]. In practical terms, this means laboratories can use synthetic datasets for research, software testing, or even public challenges without needing elaborate data use agreements. This "compliance by construction" is a major selling point of synthetic data. However, we note that ensuring *truly* zero patient re-identification risk can be tricky (addressed under challenges), so privacy advantages depend on rigorous generation and validation. Still, relative to methods like de-identification, synthetic data offers a more absolute form of privacy – since it is not derived from any one patient, it is immune to re-identification by linkage to external information [3].

4.2. Scalability and volume

Once a synthetic data generator (be it a simulation model or a neural network) is developed, it can produce large volumes of data efficiently but there are inherent limitations and costs that constrain unlimited data generation. Practical constraints include computational cost, governance controls, the need for monitoring and retraining to address population/instrument drift, and the ongoing requirement for fidelity/utility evaluation. This is transformative in fields where obtaining large datasets is expensive or slow. In the laboratory, instead of waiting years to accumulate a million patient results, one can simulate a million results overnight. Synthea's open-source platform is a prime example: it released 1,000,000 realistic synthetic patient records to the public, a volume of data that covers decades of clinical encounters and would have been nearly impossible to collect and share from real hospitals [6]. For laboratories, having access to massive data is invaluable for testing AI algorithms (which often require more data to improve) and for statistically robust validation. If one needs more cases of a rare laboratory finding one can generate as many as needed. One can also stress-test an LIS with peak loads, as an example and create millions of message transactions. This scalability also means synthetic data can be tailored to *specific rare scenarios in bulk*. For instance, one can generate 10,000 synthetic patients with adrenal tumors to evaluate a laboratory test for Cushing syndrome – far more than global case reports of that condition. In short, synthetic data removes the "small n" problem.

4.3. Flexibility and customizability

Synthetic data generation can be attuned to cover edge cases, rare conditions, or hypothetical "what-if" variations that may be absent or under-represented in real data. This flexibility is extremely useful in laboratory medicine, where unusual cases or extreme values are often the most challenging to handle. With synthetic generation, one can *intentionally model* those corner cases. For example, one can generate laboratory results for a synthetic population that has an epidemic of an uncommon disease, to evaluate how an algorithm performs under that scenario. In addition, one can adjust the distribution of patient demographics in a synthetic dataset to test if a prediction model is robust across ages and ethnicities. Synthetic data allows controlled experiments by altering parameters: one can simulate improving or worsening kidney function to evaluate how it affects a biomarker's behaviour, or simulate different instrument calibration drifts to test QC algorithms. This kind of controlled variation is difficult to extract from real-world data, which is messy and often sparse in the regions of interest. Furthermore, synthetic data can be enriched with unique combinations of features that a model could learn to recognize (e.g., an unusual pairing of laboratory results pathognomonic of a certain condition). In the deepfake histology study, the authors highlighted the ability to generate new and rare morphologies of tissues on demand [11]. Such flexibility suggests that laboratories can ensure that their tools and staff are prepared for even those one-in-a-million events.

4.4. Cost and time efficiency

Synthetic data can significantly reduce the costs associated and time spent collecting and using real clinical data. Collecting large clinical datasets often involves chart reviews, sample processing, biobanking, regulatory approvals, etc., which are both time-consuming and expensive. In contrast, once you have a good synthetic data generator, producing more data is computationally cheap. This has been noted to speed up research and development. For example, Pezoulas et al. (2024) found that many healthcare studies use synthetic data to *"reduce the cost and time required for clinical trials [especially] for rare diseases"* [7]. Synthetic patients can stand in for real ones in trial simulations, potentially reducing the number of real patients needed. Another cost benefit of using synthetic data is that it avoids the need for extensive de-identification procedures or data escrow services to handle sensitive data – processes that carry monetary and labor costs. In software testing, using synthetic data the need to maintain parallel secure environments for real patient data, again saving resources. The time efficiency is borne from the ability to iterate quickly. The simulation study by Robinson et al emphasized that analyzing certain protocols via synthetic data was possible in a short time frame, whereas doing so with real patient accrual "would not have been reproducible or repeatable" and would take years [26]. Thus, synthetic data accelerates experimentation. By some accounts, development cycles for AI models or informatics tools can be reduced because researchers do not need to wait for data access or patient recruitment – they can generate a synthetic dataset and proceed to model training and validation immediately. While generating the synthetic model in the first place is associated with significant cost, it is often amortized over many uses.

4.5. Reusability and sharing without barriers

Once created, synthetic datasets can be shared widely (given they contain no protected health information (PHI)), enabling broader collaboration. A classic limitation in laboratory medicine research is that many studies are single-institution owing to data siloing. Synthetic data offers a way to share "data experience" without sharing actual data. For example, one hospital can publish a synthetic version of its laboratory test dataset (mirroring the distributions of results and diagnoses in its region) for others to analyze. This can spur innovation in algorithms, because researchers anywhere can assess their methods on these public synthetic datasets. The *Hospitalogy* industry analysis noted that companies using synthetic health data "worry much less about HIPAA (or GDPR) compliance" – it's *"flexible and liberating"* [38]. Essentially, synthetic data can travel freely and be reused in ways real data often cannot. This multiplies its value. An algorithm tested on the synthetic data from one laboratory could then be tested on the synthetic data from another laboratory data to ensure generalizability or reproducibility, all

without any legal hurdle. In education, synthetic cases can be shared as teaching files across institutions, enriching the pool of learning materials.

4.6. Ethical and fairness benefits

Using synthetic data avoids exposing real patients to risk or burden. For instance, instead of running a potentially risky trial arm to gather data, a synthetic control arm might be used in exploratory phases (this concept is being explored in clinical trials) [39]. It also means not exposing private patient information when it is not necessary – aligning with the ethical principle of using the least sensitive data that can still answer the question. Moreover, synthetic data can be engineered to be more balanced or fair than historical real data, which might reflect societal biases. If historical data are biased (e.g., under-representing minorities or women in certain test databases), synthetic data could be generated to correct that bias and ensure AI models do not inherit it [1]. This proactive use of synthetic data for fairness is an evolving area, but it is indeed promising that one can integrate fairness by synthesizing more inclusive datasets than the raw historical record.

These advantages explain why synthetic data is often called the *"great data unlock"* in healthcare [40]– it has the potential to free us from long-standing bottlenecks. However, these benefits do not come without caveats. The next section addresses the flip side: the challenges and limitations that must be managed for synthetic data to truly deliver on its promise in the clinical laboratory.

5. Challenges and limitations

Despite its promise, synthetic data in clinical laboratory science comes with important challenges and limitations that users must be aware of (Table 3):

5.1. Fidelity to real data (validity)

A fundamental concern is whether synthetic data accurately captures the complex patterns of real clinical data. If synthetic data omits subtle correlations or rare variations present in real patients, models or conclusions derived from it may not hold up in practice [2]. Biological data are notoriously high-dimensional and can have unpredictable or unexpected relationships. For example, certain laboratory values might be correlated only in a subset of patients or only under specific physiological conditions. A synthetic data generator might miss these nuances, especially if it is trained on limited real data. The result could be synthetic patients that look plausible in isolation but collectively diverge from reality in important ways (e.g., lacking the tail of a distribution, or missing the correlation between an inflammatory marker and age in a certain disease). Evaluating fidelity is tricky – it is not enough that marginal means match; higher-order interactions matter too. Studies have highlighted the need for robust evaluation metrics for synthetic data quality [1]. For images, metrics like Frechet Inception Distance (FID) or Structural Similarity Index (SSIM) are used to quantify similarity to real images [7]. For tabular data, one examines distributions, pairwise correlations, and performance of predictive models. If fidelity is low, synthetic data could mislead. For instance, a QC algorithm validated on overly "clean" synthetic data might fail in the messy real world. Thus, ensuring data realism is a constant challenge; often it requires iterative refinement of the generative model and careful expert review of synthetic outputs.

5.2. Bias and representativeness

Synthetic data will reflect the inherent biases present in the data or assumptions used to generate it – and can even amplify them if not careful [1]. If a generative model is trained on a real dataset that is biased (say it underrepresents a demographic, or it contains predominantly normal cases with few abnormals), the synthetic data will reflect those same skewed patterns, sometimes in exaggerated form. There have been concerns that naive use of synthetic data could worsen bias in AI models. For example, if women are underrepresented in creatinine laboratory data owing to selection bias, a synthetic dataset may end up with even fewer female profiles unless explicitly corrected, leading an AI model to perform poorly for female patients. Additionally, some generative models may overfit to majority patterns and *smooth out* minorities (mode collapse in GANs is an example, where the generator might ignore rare modes of the data distribution). This can erase outliers or unique cases – ironically the opposite of what one would want in some training scenarios. Mitigating bias in synthetic data generation is an active area of research. Techniques include conditioning the generation on subgroups to ensure all are represented, or using fairness-aware algorithms that try to decorrelate sensitive attributes during generation [1]. Another approach is adversarial debiasing – adding constraints so that synthetic data cannot allow a model to easily predict protected attributes unless they are medically relevant [1]. It is also recommended to audit synthetic datasets for known biases, much like one would audit real datasets [1] In summary, synthetic data is not inherently unbiased or "magic" – it requires conscious effort to either preserve or adjust the representation of different groups and conditions as needed. Otherwise, we risk baking in the same disparities present in historical data or even creating new artifacts of bias through the generation process.

5.3. Risk of overfitting and over trusting synthetic data

If an AI model or laboratory workflow is developed using predominantly synthetic data, there is a risk that it may not generalize well to real data – essentially an overfitting to the synthetic universe. This can happen if the synthetic data, despite best efforts, fails to capture some complexities of reality. For example, an AI model may learn to detect synthetic "fingerprints" in the data (imperceptible quirks of the generator) rather than truly general patterns. One study noted the danger of models "overgeneralizing" or learning non-existent correlations when trained on synthetic data that does not accurately represent reality [2]. Such spurious correlations may be introduced by the generator or by simulation assumptions, and a model may latch onto them, performing well on synthetic validation but poorly on real data. In essence, the model mistakes the synthetic data rules for real world rules. Thus the

Table 3
Authentic vs synthetic laboratory data: key differences.

Dimension	Authentic patient data	Synthetic patient data
Primary purpose	Clinical truth; reflects real practice/outcomes	Safer sharing/experimentation; supports development/training/simulation
Privacy/confidentiality	High sensitivity; strict access control required	Lower risk but not risk-free; needs disclosure risk assessment and governance
Fidelity	Gold standard for site- and population-specific patterns	Variable; depends on method and constraints; may miss rare correlations
Bias & representativeness	Reflects existing practice and inequities	Can replicate/amplify bias; can be rebalanced—must be stated and evaluated
Shareability/collaboration	Often restricted by legal/ethical barriers	Typically easier to share; still requires approval and clear labeling
Validation role	Directly validates models/QC methods	Should not replace real-world validation; best for development/stress tests
Drift over time	Naturally captures drift over time	Must be updated/retrained and re-validated to reflect drift
Auditability/Traceability	Depends on sourcing/documentation	Requires generator provenance, versioning, and documentation

solution is to always validate models on real-world data whenever possible; synthetic data should ideally augment real data, not entirely replace it for final model tuning. Some experts have proposed hybrid training – using synthetic data for initial training or to fill gaps, but always doing a final fine-tune or calibration with a slice of real data to ensure alignment with reality. Overtrust can also extend to human users: if staff or researchers assume the synthetic data is a perfect facsimile, they might make invalid inferences. For instance, synthetic laboratory utilization data may not account for certain real-world behaviors (like seasonality, or clinicians ordering tests in bundles), therefore any conclusion drawn must be tempered with that knowledge. Essentially, synthetic data should not be viewed as ground truth but as a convenient proxy. Safeguards include gradually introducing synthetic data, checking critical metrics against known real benchmarks (e.g., does the synthetic dataset produce similar mean values, variances, and model outputs as a test real dataset?), and maintaining a healthy skepticism. This is akin to using simulation in engineering – it is powerful but always needs real-world testing.

5.4. Validation and evaluation challenges

Ensuring that synthetic data is "good enough" requires comprehensive validation protocols, which are still being standardized. Laboratories need to ask: "How do we know if our synthetic data is both useful and safe?". There are two dimensions – utility (does it maintain relevant statistical features so that analyses on synthetic data translate to real data?) and privacy (does it avoid leaking any real data information?). Achieving a balance is the privacy-utility tradeoff: overly sanitized synthetic data might be safe but not useful; highly accurate synthetic data might carry more disclosure risk [2]. Metrics for utility include comparing model performance when trained on synthetic vs authentic, as well as direct statistical similarity measures [1]. Metrics for privacy include membership inference attacks and other adversarial tests to assess if any synthetic record can be linked to a real record. A 2022 study by Zhang et al. introduced a framework for membership inference attacks against synthetic health data, which can probe if a given real patient was likely included in the training set for the generator [36]. Their results indicated that *partially* synthetic data (where some real data remains) are particularly vulnerable to such attacks, while *fully synthetic* data is structurally more resilient [36]. Nevertheless, even fully synthetic data can leak information if the generator memorized specific records (a known issue if models are overfit) [36]. Therefore, laboratories must validate privacy by simulating these attacks or using differential privacy measures. Differential privacy can be integrated to give formal guarantees, but that often comes at a cost of utility (introducing more noise into the data) [2]. To date, there is no widely agreed standard for "acceptable" privacy risk in synthetic health data – it may depend on context and risk tolerance. Another challenge is that evaluation often requires access to some real data for comparison, which is the very thing synthetic data is supposed to replace. Synthetic data is most valuable when real data is scarce, yet validating it thoroughly might need some real data. In practice, one often uses whatever data was used to train the generator as a baseline for evaluation, acknowledging that if that data was limited, some issues might escape detection.

5.5. Regulatory and acceptance hurdles

As of 2025, regulatory agencies have not fully embraced synthetic data as a substitute for real-world evidence in the contexts of device approval or clinical trial validation. Most laboratory accreditation bodies (e.g., CAP, CLIA requirements in the US) still require method validations and proficiency tests to involve actual human specimens. Regulatory bodies like the US FDA are actively researching synthetic data – the FDA has recognized the potential of generative and even published draft guidances on AI in drug development – but these stop short of allowing, say, a new diagnostic test to be approved based on synthetic data alone [5]. A recent review noted that no drug or medical device has been approved to date using predominantly synthetic data as evidence [5]. Synthetic data is seen as a supplement, not a replacement, for clinical evidence [5]. There is also a cultural acceptance issue: clinicians and auditors may be skeptical of results derived from "fake" data. One can imagine an inspector asking, "How do you know your QC procedure works for real patients if you only tested it on simulated data?" – it is a valid challenge. Thus, laboratories often must perform confirmatory tests with real specimens to satisfy requirements, even if synthetic data was used as a preliminary tool. Over time, this may change. The European Medical Agency (EMA) and other agencies are discussing how synthetic arms might be used in clinical trials or how differential privacy and synthetic data could be part of compliance solutions [5]. But as of now, conservatism prevails: synthetic data lacks the trust that real clinical data commands by default. To overcome this, more evidence and case studies are needed to show that decisions made with synthetic data are as sound as those with real data. It may also require updated guidelines. Until then, laboratories should view synthetic data as an enhancer, not a final proof, when it comes to meeting regulatory standards. Additionally, there is a risk of regulatory blind spots: synthetic data does not always clearly fall under existing definitions (Eg. Is it real-world data? Is it a derived data set?), which could be exploited in ways regulators did not anticipate [2]. For instance, unscrupulous researchers may label something synthetic to skirt data rules when it is actually just mildly perturbed real data. This calls for regulators to update definitions and perhaps certify certain synthetic data generation practices as acceptable.

5.6. Trust and buy-in from end users

Beyond regulators, the end users (clinicians, laboratory managers, patients) need to trust outputs that involved synthetic data. If an AI agent flags a critical value based on a model trained partially on synthetic data, will clinicians doubt it more? Possibly, if they are aware of it. There may be an educational gap: many will not understand what synthetic data means and might conflate it with fake or invalid data. Communicating the validation and reliability of synthetic-data-derived tools is important to gain user confidence. We have seen skepticism in some quarters, with stakeholders preferring "precise replication of original healthcare records" over synthetic approximations [2]. This is partly philosophical – a comfort with what is tangible (real data) versus something generated. Over time, as success stories accumulate (like synthetic data helping catch an analyzer flaw pre-deployment, or improving an AI agent's accuracy), this skepticism can be reduced. But it remains a hurdle: laboratories may resist using synthetic data tools if they fear unknown failure modes.

5.7. Technical expertise and resources

Generating high-quality synthetic data is not trivial. It may require specialized data science expertise and computational resources (especially for training GANs or diffusion models on large datasets). Not all laboratories have access to that. Implementing synthetic data solutions can necessitate partnerships or new hires. Additionally, maintaining synthetic data generators is an ongoing task: as real-world data drifts or new patterns emerge (e.g., a new therapy changes laboratory result distributions), the synthetic model should be updated to remain current. This continuous maintenance is a burden. There can also be significant up-front work to develop or tune a synthetic data model for data from a particular laboratory (ensuring that local idiosyncrasies like instrument biases are correctly reflected). For smaller laboratories, it might not be feasible to do this in-house. However, there are emerging platforms (some open-source, some commercial) that aim to make synthetic data generation easier – their maturity will determine how accessible this technology becomes without deep expertise.

In light of these challenges, it is clear that synthetic data is not a

panacea. It must be used with care and transparency. Best practices are still being formulated, but likely include: always document how synthetic data was generated and validated, combine synthetic with real data where possible, ensure diverse representation in generation to avoid bias, and rigorously test any model or process on real data before clinical use. In the next section, we look ahead to how the field may evolve to address these limitations and further integrate synthetic data into the fabric of laboratory medicine.

6. Future directions and opportunities

The use of synthetic data in clinical laboratories is rapidly evolving. Looking to the future, several directions appear promising for maximizing benefits while minimizing risks:

6.1. Hybrid and augmented datasets

Rather than relying solely on either real or synthetic data, future approaches will likely use *hybrid datasets* that combine the two in intelligent ways. For instance, a hybrid dataset might start with a core of real patient data and then have synthetic records added to enrich underrepresented categories [5]. This can yield a training set that has the best of both worlds: it retains real-data fidelity for common cases and injects synthetic cases to ensure coverage of rarer scenarios. As mentioned earlier, mixing synthetic and real records (with synthetic making up, say, X% of the data) can provide strong privacy (if X is large, identities are obscured) while maintaining higher accuracy on features that are well-covered by real data [5]. We anticipate methods for optimally blending synthetic with real data will be developed – for example, weighting synthetic samples less in training if they are deemed less reliable, or using synthetic data to pre-train models and real data for final fine-tuning (a strategy already being tested). Additionally, hybrid approaches could mean using synthetic data engines to insert missing data in real datasets or to perform data augmentation in targeted ways. The concept of *partially synthetic data* (replacing only certain sensitive variables with synthetic values) is also being explored as a privacy technique that still allows the real data structure to be used for the rest [5]. Careful evaluation will be needed, but hybrid strategies are likely to become standard practice.

6.2. Integration of privacy-enhancing technologies

As concerns about privacy and re-identification persist, we expect to see synthetic data generation incorporate additional privacy safeguards like differential privacy and federated learning. Differential privacy (DP) introduces statistical noise in a principled way to provably limit what can be inferred about any single individual from the output. Future synthetic data tools may provide a DP guarantee, where even if an attacker knew all but one record in the training data, they still could not confidently identify that last record from the synthetic outputs. Some initial research shows that diffusion models may inherently offer favorable privacy-utility trade-offs, and combining them with DP could further reduce leakage [12]. Another approach is federated synthetic data generation, where multiple institutions collaboratively train a generative model without sharing raw data (using federated learning), and then each can generate synthetic data from the pooled model. This way, everyone benefits from a richer model without any site exposing patient data – the model essentially becomes a vessel of combined knowledge. This could be used, for example, by a consortium of laboratories to produce a global synthetic dataset for a rare disease, which any member can then use. We also foresee better metrics and disclosure controls being standardized, possibly with regulatory input, so that laboratories can certify their synthetic data meets certain privacy criteria (analogous to de-identification certification). In the same vein, techniques like watermarking synthetic data may be developed to distinguish it from real data (to avoid accidental misuse or to trace provenance).

6.3. Regulatory evolution and standards

Regulatory bodies are increasingly engaging with the concept of synthetic data, and we expect more formal guidance to emerge. The FDA, EMA, and Medicines and Healthcare products Regulatory Agency (MHRA) have all indicated interest in how AI-generated data might support regulatory decisions [5]. For example, in 2025 the FDA released a draft guidance on using AI in drug development, and while it did not yet detail synthetic data, it is plausible that future versions will outline acceptable use cases – perhaps accepting synthetic control arms in certain clinical trials or allowing synthetic data to supplement premarket submissions for devices (with conditions on validation) [5]. The EMA reflection paper on AI already acknowledges synthetic data as a tool for augmenting training datasets and suggests aligning its use with differential privacy techniques [5]. We may see pilot programs where regulators accept synthetic data analyses as part of a package – for instance, using a synthetic dataset to demonstrate potential mitigation bias in an AI model. Additionally, standard-setting organizations such as CLSI or ISO may introduce guidelines for synthetic data in laboratory validations. Governments may also encourage synthetic data use for public health by providing synthetic versions of national health datasets for research (some have started doing so). However, until regulations concretely incorporate synthetic data, laboratories should maintain a dual strategy: use synthetic data for internal innovation, but still collect real-world evidence for official purposes. In the meantime, laboratories can contribute to shaping these policies by publishing their findings on the efficacy and limitations of synthetic datas, thereby informing regulators.

6.4. Advanced generative models and automation

On the technical front, we anticipate ever-more sophisticated generative models tailored to healthcare data. *Transformers* (popular in text and now being applied to tabular data) might lead to new synthetic data tools that handle mixed data types even better. The open-source and commercial ecosystems around synthetic data (e.g., MIT's Synthetic Data Vault, Vanderschaar Lab's SynthCity, commercial vendors like Mostly AI, Gretel, etc.) are growing and will likely become more user-friendly and validated. Automation could allow a laboratory to input its dataset and get a synthetic data generator with minimal manual tuning. Another interesting direction is combining causal or mechanistic models with generative models – for laboratory data, this could mean generative models that respect known biochemistry or physiology. For instance, a synthetic chemistry panel generator that ensures the anion gap is internally consistent, or synthetic full blood count data that respects how cell lines correlate in real bone marrow disorders. By embedding domain knowledge, synthetic data may become more reliable. Diffusion models have opened up possibilities for more controllable generation (e.g., generating data conditional on certain outcomes). So a laboratory may want to "generate 1000 synthetic patients who had outcome X" to study risk factors. The models may also incorporate real-time learning; for example, a laboratory could potentially continuously update its synthetic data model as new real data comes in (maintaining privacy by only releasing synthetic outputs).

6.5. Global collaboration and equity

One of the most exciting future directions is the role synthetic data can play in global health and equity. Laboratories in low-resource settings often lack large datasets or access to advanced training materials. Synthetic data could level the playing field by providing those laboratories with realistic datasets to train personnel, validate procedures, or develop AI tools, without needing the same volume of real data locally. As an analogy, consider how simulation training has improved medical

education in places without access to high-tech patient cases – synthetic data can achieve similarly for laboratory diagnostics. A small laboratory in a developing country could, for instance, download a synthetic dataset of thousands of malaria cases with various co-infections to train a ML model or simply to practice QC, something that could never be assembled from limited records. This fosters global collaboration because knowledge (in data form) can be shared more freely. A quote from a recent perspective resonates: *"By bridging data gaps and fostering global collaboration, synthetic data have the potential to revolutionize healthcare research"* [1]. In the laboratory context, that means enabling every laboratory, big or small, to benefit from the collective data experience of many. Organizations such as the WHO or global laboratory networks might even curate synthetic datasets for key conditions as reference standards or training sets. Moreover, synthetic data may help include populations that are often missing from research (children, certain ethnic groups, etc.) by generating plausible data for these groups based on whatever is known, thus pushing research inclusivity. While synthetic data is not a magic bullet for the lack of real data in underserved regions, it can mitigate the gap and ensure that AI and algorithms developed in data-rich environments can be tested and adapted safely for other contexts via synthetic trials.

6.6. Routine use in laboratory operations

In the future, synthetic data could become a routine part of laboratory operations. For example, continuous quality monitoring might incorporate synthetic data challenges – the LIS may periodically inject synthetic test records through the pipeline to ensure all systems (instruments, middleware, reporting) react appropriately (a form of ongoing validation). This enables detection of issues in real-time without waiting for a real patient case to expose a flaw. Laboratories might maintain digital twins of their operations fed by a stream of synthetic data that parallels real workloads, providing a constant safeguard and optimization tool. Proficiency testing programs might adopt synthetic case scenarios regularly, not just as rare pilots. Over time, success stories will build confidence: perhaps an instrument manufacturer avoids a costly recall because synthetic stress testing detected a software bug; or a laboratory averts an erroneous result release because their synthetic data sentinel triggered an alarm when a control rule failed in simulation.

In charting these future directions, it is clear that collaboration between stakeholders will be key. Data scientists, laboratory professionals, clinicians, regulators, and ethicists need to continue the dialogue on how to harness synthetic data responsibly. The technology is moving fast – what may have seemed like science fiction a decade ago (convincingly fake patient records, deepfake images indistinguishable from real) is now reality. The task ahead is to integrate these capabilities into the healthcare system in a way that enhances outcomes, maintains trust, and adheres to the primacy of patient welfare.

7. Conclusion

Synthetic data is poised to become an integral component of clinical laboratory science, offering innovative solutions to long-standing challenges of data scarcity, privacy, and system testing. It enables laboratories to design, test, and learn in silico – whether that means validating an analyzer with thousands of virtual specimens, training an AI agent on rare disease patterns, or simulating an entire laboratory workflow under stress. The examples reviewed here, from troponin assay validation with synthetic cohorts [24] to deepfake pathology images augmenting diagnosis [11] illustrate that synthetic data can closely approximate real-world scenarios and even reveal insights that would be impractical to obtain otherwise.

However, it is equally clear that synthetic data is not a panacea or a replacement for real patient data when it comes to ultimate clinical validation. The fidelity and fairness of synthetic data must be rigorously verified, and any tools developed with it must be carefully validated against real-world outcomes to ensure accuracy, generalizability, and safety [2]. In essence, synthetic data can extend what laboratories are able to do – allowing more exhaustive testing and more inclusive modeling – but it does not absolve us from connecting back to reality. As the regulatory and scientific community grapples with these new methodologies, guidelines will mature to specify how synthetic data should be used and evaluated. In the meantime, laboratories venturing into synthetic data should do so with a mindset of *augmentation and experimentation*, not total reliance.

When implemented thoughtfully, synthetic data can be a powerful tool in the laboratory informatician toolkit. It offers a safe sandbox to innovate, free from patient risk. It preserves privacy while enabling collaboration and skill-building. Finally, it holds promise for democratizing access to knowledge, so that even laboratories with limited resources can benefit from the collective data experience encoded in synthetic form [1]. The coming years will likely see synthetic data transitioning from novel research topic to a routine utility in laboratories – much like QC material and calibrators are now standard.

In summary, synthetic data in the clinical laboratory offers immense potential for training, quality assurance, AI development, and operational modeling. Its advantages – privacy, scalability, flexibility, and cost-effectiveness – make it a compelling ally for laboratory innovation. Yet, this potential will only be realized if we uphold rigorous validation, guard against bias, and maintain transparency regarding its use. With those guardrails in place, synthetic data can significantly and safely extend the capabilities of clinical laboratories, leading to improved diagnostics, more resilient systems, and ultimately better patient care.

CRediT authorship contribution statement

Tahir S. Pillay: Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Barbara S. van Deventer:** Writing – review & editing. **Siphokazi Gwiliza:** Writing – review & editing. **Evette L. Subramoney:** Writing – review & editing. **Chantal van Niekerk:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] J.M. Mendes, A. Barbar, M. Refaie, Synthetic data generation: a privacy-preserving approach to accelerate rare disease research, Front. Digital Health 7 (2025) 1563991.
[2] M. Giuffrè, D.L. Shung, Harnessing the power of synthetic data in healthcare: innovation, application, and privacy, npj Digital Med. 6 (1) (2023) 186.
[3] Moore, De-identified and unregulated: how data brokers outpace state privacy Laws, Vanderbilt J. Entertain. Technol. Law 27 (4) (2025) 863–897.
[4] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, in: D.-V. Finale, F. Jim, K. David, R. Rajesh, W. Byron, W. Jenna (Eds.), Proceedings of the 2nd Machine Learning for Healthcare Conference, PMLR, Proceedings of Machine Learning Research, 2017, pp. 286–305.
[5] G. Pasculli, M. Virgolin, P. Myles, A. Vidovszky, C. Fisher, E. Biasin, M. Mourby, F. Pappalardo, S. D'Amico, M. Torchia, A. Chebykin, V. Carbone, L. Emili, D. Roeshammar, Synthetic data in healthcare and drug development: definitions, regulatory frameworks, issues, CPT Pharmacometrics Syst. Pharmacol. 14 (5) (2025) 840–852.
[6] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. McLachlan, Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, J. Am. Med. Inform. Assoc. 25 (3) (2018) 230–238.

[7] V. Pezoulas, D. Zaridis, M. Eugenia, C. Androutsos, K. Apostolidis, N. Tachos, D. Fotiadis, Synthetic data generation methods in healthcare: a review on open-source tools and methods, Comput. Struct. Biotechnol. J. 23 (2024).
[8] A.A. Barr, J. Quan, E. Guo, E. Sezgin, Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data, Front. Artif. Intell. 8 (2025).
[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[10] M. Pozzi, S. Noei, E. Robbi, L. Cima, M. Moroni, E. Munari, E. Torresani, G. Jurman, Generating and evaluating synthetic data in digital pathology through diffusion models, Sci. Rep. 14 (1) (2024) 28435.
[11] K. Falahkheirkhah, S. Tiwari, K. Yeh, S. Gupta, L. Herrera-Hernandez, M. R. McCarthy, R.E. Jimenez, J.C. Cheville, R. Bhargava, Deepfake histologic images for enhancing digital pathology, Lab. Investig. 103 (1) (2023) 100006.
[12] A. Kotelnikov, D. Baranchuk, I. Rubachev, A. Babenko, TabDDPM: Modelling Tabular Data with Diffusion Models, Proceedings of the 40th International Conference on Machine Learning, PMLR, Honolulu, Hawaii, USA, 2023.
[13] N.Z. Petrakos, E.E.M. Moodie, N. Savy, A framework for generating realistic synthetic tabular data in a randomized controlled trial setting, Stat. Med. 44 (18–19) (2025) e70227.
[14] D. Kaur, M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, N. Markuzon, Application of Bayesian networks to generate synthetic health data, J. Am. Med. Inform. Assoc. 28 (4) (2021) 801–811.
[15] V. Lukić, S. Ignjatović, Moving average procedures as an additional tool for real-time analytical quality control: challenges and opportunities of implementation in small-volume medical laboratories, Biochem. Med. (Zagreb) 32 (1) (2022) 010705.
[16] S.A. Alajaji, Z.H. Khoury, M. Elgharib, M. Saeed, A.R.H. Ahmed, M.B. Khan, T. Tavares, M. Jessri, A.C. Puche, H. Hoorfar, I. Stojanov, J.J. Sciubba, A.S. Sultan, Generative adversarial networks in digital histopathology: current applications, limitations, ethical considerations, and future directions, Mod. Pathol. 37 (1) (2024) 100369.
[17] G. Grazhdanski, V. Vasilev, S. Vassileva, D. Taskov, I. Antova, I. Koychev, S. Boytcheva, SynthMedic: utilizing large language models for synthetic discharge summary generation, correction and validation, J. Biomed. Inform. 170 (2025) 104906.
[18] O. Ferrández, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, BoB, a best-of-breed automated text de-identification system for VHA clinical documents, J. Am. Med. Inform. Assoc. 20 (1) (2012) 77–83.
[19] R. Zhou, Y.-F. Liang, H.-L. Cheng, W. Wang, D.-W. Huang, Z. Wang, X. Feng, Z.-W. Han, B. Song, A. Padoan, M. Plebani, Q.-T. Wang, A highly accurate delta check method using deep learning for detection of sample mix-up in the clinical laboratory, Clinical Chemistry and Laboratory Medicine (CCLM) 60 (12) (2022) 1984–1992.
[20] S. Sewpersad, B. Chale-Matsau, T.S. Pillay, Real world feasibility of patient-based real time quality control (PBRTQC) using five analytes in a South African laboratory, Clin. Chim. Acta 565 (2025) 120006.
[21] N. Lorde, S. Mahapatra, T. Kalaria, Machine learning for patient-based real-time quality control (PBRTQC), analytical and preanalytical error detection in clinical laboratory, Diagnostics (Basel) 14 (16) (2024).
[22] X. Duan, M. Zhang, Y. Liu, W. Zheng, C.Y. Lim, S. Kim, T.P. Loh, W. Guo, R. Zhou, T. Badrick, Next-generation patient-based real-time quality control models, Ann. Lab. Med. 44 (5) (2024) 385–391.
[23] M.A. Cervinski, A. Bietenbeck, A. Katayev, T.P. Loh, H.H. van Rossum, T. Badrick, Advances in clinical chemistry patient-based real-time quality control (PBRTQC), Adv. Clin. Chem. 117 (2023) 223–261.
[24] J.W. Pickering, J.M. Young, P.M. George, A.S. Watson, S.J. Aldous, T. Verryt, R. W. Troughton, C.J. Pemberton, A.M. Richards, L.A. Cullen, F.S. Apple, M.P. Than, Derivation and validation of thresholds using synthetic data methods for single-test screening of emergency department patients with possible acute myocardial infarction using a point-of-care troponin assay, J. Appl. Lab. Med. 9 (3) (2024) 526–539.
[25] A. Prasanna, B. Jing, G. Plopper, K.K. Miller, J. Sanjak, A. Feng, S. Prezek, E. Vidyaprakash, V. Thovarai, E.J. Maier, A. Bhattacharya, L. Naaman, H. Stephens, S. Watford, W.J. Boscardin, E. Johanson, A. Lienau, Synthetic health data can augment community research efforts to better inform the public during emerging pandemics, medRxiv (2023) 23298687, https://doi.org/10.1101/2023.12.11.23298687.
[26] S. Robinson, F. FitzGibbon, J. Eatock, T. Hunniford, D. Dixon, B.J. Meenan, Application of synthetic patient data in the assessment of rapid rule-out protocols using point-of-care testing during chest pain diagnosis in a UK emergency department, J. Simulat. 3 (2009) 163–170.
[27] N.I.-H. Kuo, O. Perez-Concha, M. Hanly, E. Mnatzaganian, B. Hao, M. Di Sipio, G. Yu, J. Vanjara, I.C. Valerie, J. de Oliveira Costa, T. Churches, S. Lujic, J. Hegarty, L. Jorm, S. Barbieri, Enriching data science and health care education: application and impact of synthetic data sets through the health gym project, JMIR Med. Educ. 10 (2024) e51388.
[28] C. Yan, Z. Zhang, S. Nyemba, Z. Li, Generating synthetic electronic health record data using generative adversarial networks: tutorial, JMIR AI 3 (2024) e52615.
[29] H. Rashidi, AI 201: Introduction to Synthetic Data and Generative AI. https://www.cap.org/calendar/webinars/ai-201-introduction-to-synthetic-data-and-generative-ai, 2025.
[30] R.E. Foraker, S.C. Yu, A. Gupta, A.P. Michelson, J.A. Pineda Soto, R. Colvin, F. Loh, M.H. Kollef, T. Maddox, B. Evanoff, H. Dror, N. Zamstein, A.M. Lai, P.R.O. Payne, Spot the difference: comparing results of analyses from real patient data and synthetic derivatives, JAMIA Open 3 (4) (2020) 557–566.
[31] J. Pantanowitz, C.D. Manko, L. Pantanowitz, H.H. Rashidi, Synthetic data and its utility in pathology and laboratory medicine, Lab. Investig. 104 (8) (2024) 102095.
[32] S. D'amico, D. Dall'Olio, C. Sala, L. Dall'Olio, E. Sauta, M. Zampini, G. Asti, L. Lanino, G. Maggioni, A. Campagna, Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology, JCO Clinical Cancer Informatics 7 (2023) e2300021.
[33] S. Wharrie, Z. Yang, V. Raj, R. Monti, R. Gupta, Y. Wang, A. Martin, L.J. O'Connor, S. Kaski, P. Marttinen, P.F. Palamara, C. Lippert, A. Ganna, HAPNEST: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes, Bioinformatics 39 (9) (2023).
[34] F. MacKenzie, R. Marrington, Comparisons of real versus synthetic proficiency testing items, Accred. Qual. Assur. 29 (5) (2024) 333–343.
[35] K. El Emam, L. Mosquera, J. Bass, Evaluating identity disclosure risk in fully synthetic health data: model development and validation, J. Med. Internet Res. 22 (11) (2020) e23139.
[36] Z. Zhang, C. Yan, B.A. Malin, Membership inference attacks against synthetic health data, J. Biomed. Inform. 125 (C) (2022) 12.
[37] L. Pilgram, H. Ko, A. Tung, K. El Emam, Protecting patient privacy in tabular synthetic health data: a regulatory perspective, NPJ Digit Med. 8 (1) (2025) 732.
[38] AI Mostly. The Potential of Synthetic Data, 2024. https://mostly.ai/blog/potential-of-synthetic-data.
[39] M. Delleani, Synthetic data for clinical research and innovation: opportunities, challenges and future directions, ESMO Real World Data and Digital Oncology 10 (2025) 100651, https://doi.org/10.1016/j.esmorw.2025.100651.
[40] B. Madden. Synthetic Data in Healthcare: The Great Data Unlock, 2023. https://hospitalogy.com/articles/2023-11-02/synthetic-data-in-healthcare-great-data-unlock.