ORIGINAL RESEARCH **Open access**

Generative Adversarial Networks for Synthetic Longitudinal EHR Data of Type 2 Diabetes: A Framework for Preserving Temporal Treatment Effects and Comorbidity Patterns

Hiroshi Nakamura[1*], Yuta Kato[1]

Abstract

Type 2 diabetes affects over 400 million people worldwide and requires lifelong management through continuous monitoring of laboratory values, medications, and comorbidities, yet the use of longitudinal electronic health records for research is restricted by privacy regulations such as HIPAA and GDPR, making synthetic data generation an important alternative for preserving utility while protecting confidentiality. However, existing synthetic data models often fail to accurately capture temporal treatment effects and the gradual development of comorbidities, limiting their usefulness for downstream clinical and machine learning applications. To address this, a time-series generative adversarial network is proposed for longitudinal diabetes data, incorporating a temporal encoder for irregular sampling, a treatment-conditioned generator, and dual discriminators that evaluate both static patient characteristics and dynamic clinical trajectories to ensure consistency between interventions and outcomes. By explicitly modeling temporal dependencies and comorbidity structures, the framework produces more realistic synthetic patient records that better reflect disease progression and medication-response relationships, thereby enabling privacy-preserving data sharing while supporting robust secondary analyses and future applications in chronic disease modeling.

Keywords Generative adversarial networks, Synthetic electronic health records, Type 2 diabetes, Longitudinal data, Temporal treatment effects, Comorbidity patterns

*Correspondence:
Hiroshi Nakamura
hiroshi.nakamura@outlook.com
[1] Department of AI in Healthcare Engineering, Graduate School of Medicine, Nagoya University, Nagoya, Japan

Introduction

Type 2 diabetes is a chronic disease requiring lifelong management, and longitudinal electronic health record data capture medication changes, laboratory results such as HbA1c and glucose levels, and the gradual development of comorbidities including nephropathy, retinopathy, and cardiovascular disease [1, 2]. These data streams document the interplay between therapeutic interventions and clinical progression over multiple years, providing an essential resource for understanding disease dynamics. Privacy regulations nevertheless restrict open access, underscoring the need for alternative data sources that retain analytical utility.

Privacy regulations such as HIPAA and GDPR impose strict limits on the sharing of real electronic health record data for research purposes. Synthetic data generation therefore emerges as a practical solution that can expand access while reducing re-identification risks. Existing methods, however, often generate outputs that fail to preserve the temporal dynamics inherent in chronic disease trajectories [3, 4].

Standard generative adversarial networks were originally developed for static data and produce outputs that lack sequential dependencies. Time-series variants have advanced the field yet still encounter difficulties in modeling treatment effects such as the lagged response to metformin initiation on subsequent HbA1c trajectories. Comorbidity patterns likewise remain challenging to replicate faithfully across extended time horizons [5, 6].

This paper proposes a conceptual framework for a time-series generative adversarial network specifically designed for type 2 diabetes longitudinal electronic health record data. The framework explicitly models temporal treatment effects, including medication changes and their downstream clinical consequences, while preserving comorbidity patterns. Key innovations comprise treatment-conditioned generation and dual discriminators that jointly evaluate static and temporal fidelity; subsequent sections detail the architecture, data representation strategies, and design principles that underpin this approach [7].

Background

Type 2 diabetes longitudinal data

Longitudinal electronic health record data for type 2 diabetes typically encompass static demographic variables, time-varying laboratory measurements, medication dispensing records, and diagnostic codes that accumulate over years of follow-up. These elements collectively describe individualized trajectories of glycemic control, treatment escalation, and the sequential onset of microvascular and macrovascular complications. Capturing such multilayered temporal information is essential for any generative model that seeks clinical realism [1].

The progression of type 2 diabetes is marked by predictable yet patient-specific patterns, including gradual rises in HbA1c, stepwise intensification of oral agents or insulin, and the delayed emergence of comorbidities such as chronic kidney disease or peripheral neuropathy. These patterns reflect both biological mechanisms and clinical decision-making processes documented in routine care. A generative framework must therefore encode the temporal ordering and conditional dependencies that link treatment events to later outcomes [8].

Synthetic data in healthcare

Synthetic electronic health record data have gained prominence as a means to support algorithm development, facilitate data sharing for multi-center studies, and enable educational simulations without exposing protected health information. Their privacy advantages stem from the absence of direct linkage to real individuals, thereby circumventing many regulatory barriers that constrain conventional data use. Nonetheless, the generated records must demonstrate sufficient utility to justify replacement of real data in downstream tasks [9].

Utility concerns arise when synthetic datasets deviate from the statistical properties or causal structures of real-world records, potentially leading to biased models or erroneous inferences. Prior work has emphasized the importance of maintaining both marginal distributions and joint temporal relationships across patient sequences. The conceptual framework presented here addresses these concerns by prioritizing fidelity to treatment effects and comorbidity evolution within type 2 diabetes cohorts [10].

Time-series GANs

Yoon *et al*. introduced TimeGAN, a framework specifically designed for generating realistic time-series data while preserving temporal dynamics through a combination of supervised and unsupervised objectives [5]. Subsequent extensions such as C-RNN-GAN and DoppelGANger have adapted recurrent architectures to handle variable-length sequences commonly found in clinical settings. These models demonstrate the feasibility of synthesizing sequential observations yet reveal limitations when applied directly to electronic health record streams that embed causal treatment relationships.

Table 1 contrasts the theoretical limitations of conventional time-series GANs with the proposed framework, highlighting key architectural innovations that address clinical and temporal deficiencies

Table 1. Comparative Theoretical Limitations of Standard Time-Series GANs versus the Proposed Treatment-Aware Framework

Dimension	Standard Time-Series GANs	Proposed Framework	Theoretica Advanceme
Treatment Effect	Implicit or absent	Explicit conditioning	Enables causal

Modeling		on treatment history	interpretabil
Temporal Ordering	Learned implicitly	Enforced via constraints and loss functions	Prevents reversed causality
Irregular Sampling Handling	Often ignored or interpolated crudely	Explicit temporal encoding with time gaps	Preserves real-world observatio structure
Comorbidity Modeling	Marginal or pairwise only	Higher-order correlation preservation	Captures complex disease interaction
Static vs Temporal Evaluation	Single discriminator	Dual discriminators (static + temporal)	Improves multi-leve fidelity
Clinical Plausibility	Weak domain integration	Embedded domain knowledge and guidelines	Aligns outpu with real ca pathways
Counterfactual Validity	Limited	Treatment-consistent trajectory generation	Supports causal inference studies
Long-Term Dependency Capture	Limited stability	Designed for chronic disease progression	Improves longitudina coherence

Time-series generative adversarial networks exhibit strengths in reproducing autocorrelation structures and marginal distributions yet frequently overlook domain-specific constraints such as the precedence of medication changes before observable clinical responses. In the context of chronic disease management, these omissions can distort the simulated impact of therapeutic interventions on long-term outcomes. The proposed framework therefore augments core time-series GAN components with explicit conditioning and dual-discriminator mechanisms to better align generated data with clinical realities [11].

Framework Overview

High-level architecture

The high-level architecture begins with real longitudinal electronic health record sequences that undergo processing through a temporal encoder responsible for embedding irregularly sampled observations into a continuous latent representation. The generator then receives this encoded context together with random noise to produce synthetic patient trajectories that mirror the original data distribution. Dual discriminators subsequently evaluate the synthetic outputs, one focusing on static patient-level features and the other on the fidelity of temporal dynamics [7].

Figure 1 illustrates the hierarchical architecture of the proposed treatment-conditioned time-series GAN framework, showing how longitudinal EHR data are encoded, conditioned on treatment history, and processed through dual discriminators to generate clinically coherent synthetic type 2 diabetes trajectories.

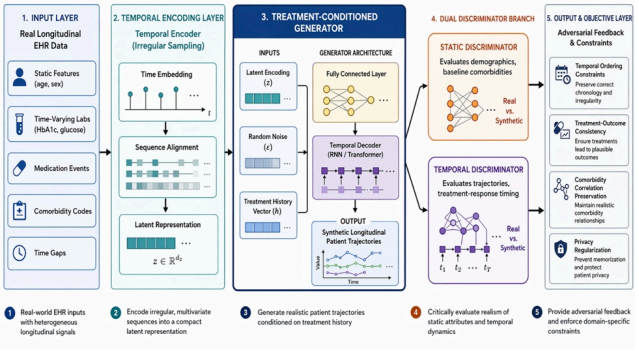

Figure 1. Hierarchical Architecture of a Treatment-Conditioned Time-Series GAN for Synthetic Longitudinal Type 2 Diabetes EHR Data

This modular design ensures that information flows from observed clinical events through latent conditioning signals into generated sequences while maintaining separation between static and time-dependent evaluations. The encoder compresses heterogeneous data types into unified embeddings, allowing the generator to reconstruct plausible future states conditioned on past treatment histories. Such an architecture supports end-to-end training that balances realism against privacy objectives [3].

Core assumptions

The framework assumes that input electronic health record data are structured with consistent time stamps, medication and diagnosis coding standards, and sufficient longitudinal

depth to reveal treatment-response patterns. These assumptions align with contemporary clinical databases that record encounters, prescriptions, and laboratory results using standardized terminologies. Under these conditions, the generative process can learn meaningful temporal mappings without requiring manual feature engineering [9].

Additional assumptions include the availability of large-scale de-identified cohorts containing thousands of patient-years of follow-up, which provide the statistical power necessary for adversarial training. Medication exposure and comorbidity indicators are treated as observable events whose ordering must be preserved in synthetic outputs. The framework therefore operates under the premise that causal relationships observed in real data can be approximated through conditional generation mechanisms [1].

Design principles

Design principles emphasize explicit preservation of temporal treatment effects by conditioning generation steps on prior medication events and enforcing causal precedence within each synthetic sequence. Comorbidity pattern fidelity is achieved through correlation-aware loss terms that maintain both pairwise associations and higher-order temporal trajectories. Privacy guarantees are integrated by design through differential privacy considerations and distance-based identifiability controls [10].

These principles collectively guide architectural choices toward clinical plausibility rather than purely statistical matching. Treatment effect preservation ensures that simulated medication initiations precede corresponding changes in laboratory values, while comorbidity constraints prevent implausible co-occurrences. The resulting framework thereby produces synthetic data that remain analytically useful for research on diabetes progression and management [12].

Table 2 analytically decomposes the framework into its core components, clarifying how each module contributes to temporal fidelity, causal consistency, and clinical realism

Table 2. Analytical Decomposition of Framework Components and Their Functional Roles in Preserving Temporal and Clinical Fidelity

Component	Input Dependencies	Core Mechanism	Tem R
Temporal Encoder	Irregular time-series EHR data	Time-aware embedding with latent compression	Aligns i sampli conti repres
Treatment-Conditioned Generator	Latent encoding + noise + treatment history	Conditional sequence generation via RNN/Transformer	Propa treat effects in t
Temporal Decoder	Latent states and prior outputs	Autoregressive sequence generation	Mair sequ deper acros ste
Static Discriminator	Aggregated patient features	Distributional comparison	Non-te (glo evalu
Temporal Discriminator	Full patient sequences	Sequence-level adversarial evaluation	Det tem inconsi
Constraint Mechanisms	Generated sequences + domain rules	Loss penalties and priors	Enfo orderi la consi
Comorbidity Correlation Module	Multivariate disease patterns	Correlation-preserving regularization	Mair cross depen

Temporal Data Representation

Feature types and encoding

Feature types within type 2 diabetes longitudinal records include static demographic and baseline comorbidity variables, time-varying laboratory and vital-sign measurements, and discrete event indicators for medication starts, stops, or dosage changes. Encoding strategies convert these heterogeneous inputs into fixed-

dimensional embeddings suitable for sequential processing, employing one-hot representations for categorical events and normalized continuous scales for laboratory values. The resulting unified representation captures both instantaneous states and cumulative exposure histories [13].

Static features such as age, sex, and initial comorbidity burden anchor the patient profile, whereas time-varying elements encode evolving clinical status. Event-based tokens explicitly mark treatment transitions, allowing the model to learn conditional dependencies between interventions and subsequent observations. This multi-type encoding scheme therefore forms the foundation for subsequent generator and discriminator modules [4].

Handling irregular sampling

Clinical visits in type 2 diabetes care occur at irregular intervals dictated by patient adherence, disease severity, and scheduling constraints, necessitating time-aware representations that explicitly incorporate elapsed time between observations. Masking techniques or interpolation layers can bridge gaps while preserving the original temporal structure, ensuring that generated sequences respect the same irregularity patterns observed in real data. The framework therefore treats time deltas as additional input features rather than assuming uniform sampling [7].

Irregular sampling introduces challenges for sequence modeling because standard recurrent architectures expect fixed intervals. By embedding absolute or relative timestamps alongside clinical variables, the temporal encoder learns to modulate hidden states according to actual observation density. This approach maintains fidelity to the sporadic nature of routine diabetes monitoring and prevents artificial smoothing that could distort treatment effect estimation [11].

Generator Architecture

Latent space and conditioning

The generator operates within a latent space that combines random noise vectors with an explicit treatment conditioning vector derived from prior medication sequences. Conditioning on treatment histories enables the model to produce trajectories that reflect realistic responses to specific antidiabetic regimens, such as the expected decline in HbA1c following metformin initiation. This mechanism ensures that synthetic data respect the causal directionality documented in clinical practice [5].

Random noise provides stochastic variation across patients while the conditioning vector injects deterministic domain knowledge about medication effects. The combined input is passed through an initial fully connected layer before entering recurrent or transformer-based temporal decoding blocks. Consequently, each generated sequence begins with a treatment-informed context that propagates forward in time [3].

Temporal decoder

The temporal decoder employs recurrent neural network or transformer layers to autoregressively generate sequential outputs comprising laboratory values, medication indicators, and diagnosis flags at each time step. These layers maintain hidden states that accumulate information from previous steps, allowing the model to produce coherent long-range trajectories consistent with type 2 diabetes progression. Output heads are specialized for different data modalities to accommodate mixed continuous and discrete variables [11].

Autoregressive generation proceeds step-by-step, with each prediction conditioned on both the latent context and the previously generated tokens. This design replicates the cumulative nature of chronic disease records in which current laboratory results depend on recent treatment adjustments. The decoder therefore produces complete patient timelines that exhibit plausible temporal evolution from baseline through extended follow-up [7].

Treatment effect modeling

Treatment effect modeling within the generator incorporates explicit causal constraints that enforce the temporal precedence of medication events before corresponding clinical outcome changes. By embedding known pharmacological lags as soft priors, the architecture discourages implausible sequences in which laboratory improvements precede rather than follow treatment initiation. Such constraints enhance the clinical interpretability of synthetic trajectories [12].

Domain-specific knowledge is further integrated through auxiliary loss terms that reward consistency with established treatment-response relationships observed in diabetes literature. The generator learns to produce counterfactual-consistent sequences that maintain internal

validity across simulated intervention scenarios. This treatment-aware generation paradigm distinguishes the framework from generic time-series models and directly addresses the preservation of causal temporal effects central to type 2 diabetes research [8].

Discriminator Architecture

Static discriminator

The static discriminator component evaluates global patient characteristics including age, sex, baseline laboratory measurements, and the overall count of comorbidities to ensure that the synthetic cohort matches the population-level statistics observed in real type 2 diabetes data. This evaluation occurs independently of the temporal sequence and focuses on the fidelity of static attributes that define patient subgroups. By comparing synthetic and real distributions of these features, the discriminator enforces demographic and baseline clinical balance across generated records. Additional checks on comorbidity prevalence further guarantee that the synthetic data reflect the expected disease burden in diabetic populations [14].

Operating on summary statistics extracted from the full patient trajectory, the static discriminator provides a holistic assessment that prevents mode collapse in static dimensions. It contributes to overall model stability by penalizing deviations in key demographic and comorbidity aggregates. This design choice draws from established practices in synthetic health data generation where global fidelity is prioritized alongside sequential realism. The result is a more representative synthetic population suitable for epidemiological analyses [15].

Temporal discriminator

The temporal discriminator specifically assesses the dynamics of sequences such as HbA1c trajectories and the timing of medication responses within each synthetic patient record. It employs recurrent or convolutional layers to process the entire time series and detect inconsistencies in progression patterns or treatment effect lags. By focusing on sequential dependencies, this module ensures that generated data preserve the autocorrelation and cross-correlation structures inherent to longitudinal diabetes monitoring. Such scrutiny is critical for maintaining the clinical plausibility of evolving laboratory values and event timings [16].

Utilizing architectures capable of capturing long-range dependencies, the temporal discriminator differentiates real from synthetic sequences based on their dynamic properties rather than isolated snapshots. It reinforces the generator's ability to produce coherent trajectories that align with observed disease progression rates. Integration of this component within the adversarial framework enhances the model's sensitivity to temporal irregularities typical in electronic health records. Overall, it safeguards the fidelity of time-dependent relationships essential for downstream predictive modeling in type 2 diabetes [17].

Preserving Treatment Effects

Treatment-outcome consistency

Treatment-outcome consistency is maintained by ensuring that synthetic data accurately reflect established clinical relationships between medication initiations and subsequent changes in glycemic control. For instance, the framework enforces that reductions in HbA1c follow rather than precede the start of therapies like metformin or insulin. This consistency supports the validity of counterfactual analyses performed on the generated datasets. Counterfactual consistency further requires that alternative treatment paths lead to plausible outcome shifts consistent with pharmacological knowledge [18].

The generator incorporates mechanisms to verify that observed improvements or deteriorations in clinical markers align temporally with documented treatment adjustments. Such alignment prevents the creation of unrealistic scenarios that could mislead researchers studying comparative effectiveness. By embedding these checks, the framework enhances the reliability of synthetic data for causal inference tasks. This approach addresses a key limitation in prior generative models applied to chronic disease data [19].

Temporal ordering constraints

Temporal ordering constraints require that treatment events always precede their associated clinical outcomes within each generated sequence. The model penalizes any violation where laboratory improvements appear before the corresponding medication change, thereby enforcing logical causality. These constraints are implemented through specialized loss functions that monitor the relative positioning of events and responses. Such ordering is

fundamental to replicating the decision-making processes observed in real-world diabetes management [20].

By explicitly modeling the precedence of interventions over outcomes, the framework avoids common artifacts in time-series generation where sequences exhibit reversed causality. This mechanism strengthens the internal validity of synthetic trajectories for longitudinal studies. It also facilitates more accurate simulations of treatment escalation patterns over extended follow-up periods. The constraints thus contribute to the overall temporal integrity of the synthetic electronic health records [21].

Domain knowledge integration

Domain knowledge integration occurs through the incorporation of clinical guidelines as soft constraints within the generator's objective function. These guidelines inform reward mechanisms that favor trajectories consistent with established standards of care for type 2 diabetes. For example, the model rewards sequences that demonstrate appropriate medication intensification in response to persistent hyperglycemia. This integration ensures that synthetic data align with expert-derived expectations of disease management [22].

Soft constraints derived from diabetes literature guide the generation process without overly restricting the diversity of patient-specific responses. The resulting reward functions promote clinically plausible pathways while allowing for natural variation across individuals. Such knowledge-driven elements elevate the framework beyond purely data-driven approaches common in generic GANs. Consequently, the synthetic records become more interpretable and actionable for healthcare researchers [23].

Preserving Comorbidity Patterns

Comorbidity trajectories

Comorbidity trajectories in type 2 diabetes are characterized by the progressive development of complications such as nephropathy, neuropathy, retinopathy, and cardiovascular disease over extended time periods. The framework preserves the temporal order of comorbidity onset by conditioning the generator on prior disease markers and enforcing realistic progression rates. This ensures that synthetic sequences exhibit the expected delays between initial diabetes diagnosis and subsequent complication emergence. Accurate modeling of these trajectories is vital for studies examining long-term complication risks [24].

By capturing the ordered appearance of multiple comorbidities, the model avoids generating implausible simultaneous onsets that do not reflect real patient histories. The design incorporates mechanisms to simulate the cumulative burden of complications as time advances. Such fidelity supports research into the interplay between glycemic control and complication incidence. The approach thereby enhances the utility of synthetic data for prognostic modeling in chronic care [25].

Correlation structure

The correlation structure among comorbidities is preserved through mechanisms that maintain both pairwise associations and higher-order interactions observed in real type 2 diabetes cohorts. Static correlations ensure that patients with certain baseline profiles exhibit consistent comorbidity clusters in synthetic outputs. Temporal correlations further link the evolution of one complication to the likelihood of others developing later. These structures are enforced via dedicated regularization terms in the adversarial training process [26].

Higher-order correlations capture complex dependencies such as the joint progression of cardiovascular and renal complications under poor glycemic control. The framework evaluates these relationships at both population and individual trajectory levels to prevent distortion. Preservation of the full correlation matrix contributes to the multivariate realism of generated records. This comprehensive approach distinguishes the model from simpler generative techniques that address only marginal distributions [27].

Privacy and Utility Evaluation

Privacy metrics

Privacy metrics focus on susceptibility to membership inference attacks, the distance to the closest real record, and overall identifiability risk to quantify the protection offered by the synthetic dataset. Differential privacy budgets can be incorporated during training to provide formal guarantees against re-identification. These evaluations ensure that individual patient information cannot be reverse-engineered from the generated records.

The framework prioritizes these safeguards to comply with stringent healthcare data regulations [28].

Distance-based measures such as the nearest-neighbor distance help assess how closely synthetic samples approximate the original data manifold without direct overlap. Identifiability risk assessments further validate that no synthetic record can be uniquely linked back to a real patient. By systematically applying these metrics, the model balances privacy preservation with data utility. This dual consideration is essential for the responsible deployment of synthetic electronic health records in research settings [5].

Utility evaluation

Utility evaluation examines the performance of downstream tasks such as type 2 diabetes progression prediction and treatment effect estimation when models are trained on synthetic versus real data. Temporal dynamics metrics including autocorrelation and cross-correlation functions provide additional benchmarks for sequence fidelity. These assessments confirm that synthetic datasets support equivalent analytical conclusions to their real counterparts. Comparative analyses highlight the framework's ability to retain predictive power across multiple clinical endpoints [29].

By comparing real and synthetic distributions in task-specific contexts, the evaluation quantifies how well the generated data serve as proxies for privacy-sensitive originals. Downstream task performance serves as the ultimate test of utility, ensuring that synthetic records enable reliable algorithm development and hypothesis testing. The framework's design choices directly contribute to high utility scores by prioritizing clinically relevant temporal and comorbidity features. Such rigorous evaluation protocols establish confidence in the synthetic data for real-world healthcare applications [3].

Conclusion

The proposed framework introduces a time-series generative adversarial network tailored for type 2 diabetes longitudinal electronic health record data that successfully preserves temporal treatment effects and comorbidity patterns. By integrating specialized conditioning and dual-discriminator components, the architecture generates synthetic sequences that mirror the causal and progressive nature of real patient records. This conceptual design overcomes limitations of generic models by embedding domain-specific constraints throughout the generation process. The result is a robust solution for producing privacy-compliant datasets that retain full analytical value for diabetes research.

Key innovations include the treatment conditioning mechanism that links medication sequences directly to clinical outcomes, the dual discriminators that separately validate static and temporal fidelity, and the temporal ordering constraints that enforce causal precedence. These elements work synergistically to create synthetic data that are both statistically faithful and clinically interpretable. The treatment-aware generator and comorbidity-aware correlation terms represent significant advancements over standard time-series GAN approaches. Collectively, these innovations enable the framework to address the unique challenges posed by chronic disease longitudinal data.

Limitations of the framework include the requirement for large training datasets to achieve stable adversarial convergence, the substantial computational cost associated with training recurrent or transformer-based components, and the need for extensive external validation before widespread adoption. These factors may limit immediate applicability in resource-constrained environments or smaller cohorts. Additionally, the conceptual nature of the design necessitates careful hyperparameter tuning and sensitivity analyses in future implementations. Despite these challenges, the framework provides a solid foundation for advancing synthetic data methodologies in healthcare.

Future work should focus on implementation and benchmarking of the framework using established public type 2 diabetes cohorts such as SUPREME-DM, OPTUM, and MIMIC-IV to demonstrate practical feasibility. Comparative evaluations against standard time-series GANs will further quantify the benefits of the proposed treatment-effect and comorbidity-preserving mechanisms. Such empirical validation will guide refinements and promote broader adoption within the artificial intelligence for healthcare community. Ultimately, successful deployment will facilitate secure, scalable data sharing that accelerates research into diabetes prevention and management strategies.

Acknowledgements

None

Conflict of interest

None

Ethics statement

None

Financial support

None

Received: 15 Feb 2022 Revised: 31 May 2022 Accepted: 12 Jul 2022
Published online: 20 January 2023

References

Manzini E, Vlacho B, Franch-Nadal J, Escudero J, Génova A, Reixach E, et al. Longitudinal deep learning clustering of Type 2 Diabetes Mellitus trajectories using routinely collected health records. J Biomed Inform. 2022;135:104218. https://doi.org/10.1016/j.jbi.2022.104218.

Lavikainen P, Aarnio E, Linna M, Jalkanen K. Data-driven long-term glycaemic control trajectories and their associated health and economic outcomes in Finnish patients with incident type 2 diabetes. PLoS One. 2022;17(6):e0269245. https://doi.org/10.1371/journal.pone.0269245.

Baowaly MK, Lin CC, Liu CL, Chen KT. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc. 2019;26(3):228-41.

Rashidian S, Wang F, Moffitt R, Garcia V, Dutt A, Chang W, et al. SMOOTH-GAN: towards sharp and smooth synthetic EHR data generation. In: Artif Intell Med. 2020. p. 37-48. https://doi.org/10.1007/978-3-030-59137-3_4.

Yoon J, Jarrett D, Van der Schaar M. Time-series generative adversarial networks. Adv Neural Inf Process Syst. 2019;32:5508-18.

Ghosheh G, Li J, Zhu T. A review of Generative Adversarial Networks for Electronic Health Records: applications, evaluation measures and data sources. arXiv [Preprint]. 2022:arXiv:2203.07018. https://doi.org/10.48550/arXiv.2203.07018.

Zhang Z, Yan C, Lasko TA, Sun J, Malin BA. SynTEG: a framework for temporal structured electronic health data simulation. J Am Med Inform Assoc. 2021;28(3):596-604.

Dworzynski P, Aasbrenn M, Rostgaard K, Melbye M, Gerds TA, Hjalgrim H, et al. Nationwide prediction of type 2 diabetes comorbidities. Sci Rep. 2020;10(1):1776. https://doi.org/10.1038/s41598-020-58673-y.

Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: a systematic review. Neurocomputing. 2022;493:28-45. https://doi.org/10.1016/j.neucom.2022.04.043.

Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett KP. Generation and evaluation of privacy preserving synthetic health data. Neurocomputing. 2020;416:244-55. https://doi.org/10.1016/j.neucom.2020.07.070.

Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional GANs. arXiv [Preprint]. 2017:arXiv:1706.02633. https://doi.org/10.48550/arXiv.1706.02633.

Zhang Z, Yan C, Malin BA. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. J Am Med Inform Assoc. 2022;29(11):1890-8.

Kirk IK, Simon C, Banasik K, Holm PC, Haue AD, Jensen PB, et al. Linking glycemic dysregulation in diabetes to symptoms, comorbidities, and genetics through EHR data mining. Elife. 2019;8:e44941. https://doi.org/10.7554/eLife.44941.

Dash S, Yale A, Guyon I, Bennett KP. Medical time-series data generation using generative adversarial networks. In: Artif Intell Med. 2020. p. 382-91. https://doi.org/10.1007/978-3-030-59137-3_35.

Gupta M, Phan TL, Bunnell HT, Beheshti R. Concurrent imputation and prediction on EHR data using bi-directional GANs: Bi-GANs for EHR imputation and prediction. In: Proc 12th ACM Int Conf Bioinformatics Comput Biol Health Inform. 2021. p. 1-9. https://doi.org/10.1145/3459930.3469521.

Sun S, Wang F, Rashidian S, Kurc T, Abell-Hart K, Hajagos J, et al. Generating longitudinal synthetic EHR data with recurrent autoencoders and generative adversarial networks. In: Data Management and Analytics for Medicine and Healthcare. 2021. p. 153-65. https://doi.org/10.1007/978-3-030-93653-2_14.

Venugopal R, Shafqat N, Venugopal I, Tillbury BM, Stafford HD, Bourazeri A. Privacy preserving generative adversarial networks to model electronic health records. Neural Netw. 2022;153:339-48. https://doi.org/10.1016/j.neunet.2022.06.020.

Torfi A, Fox EA, Reddy CK. Differentially private synthetic medical data generation using convolutional GANs. Inf Sci. 2022;586:485-500. https://doi.org/10.1016/j.ins.2021.11.055.

Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett KP. Assessing privacy and quality of synthetic health data. In: Proc Conf Artif Intell Data Discov Reuse. 2019. p. 1-4. https://doi.org/10.1145/3313294.3313388.

Xiang X, Duan S, Pan H, Han P, Cao J, Liu C. From one-hot encoding to privacy-preserving synthetic electronic health records embedding. In: Proc Int Conf Cyberspace Innov Adv Technol. 2020. p. 407-13. https://doi.org/10.1145/3444370.3444586.

Ehrhart M, Resch B, Havas C, Niederseer D. A conditional GAN for generating time series data for stress detection in wearable physiological sensor data. Sensors (Basel). 2022;22(16):5969. https://doi.org/10.3390/s22165969.

Brophy E, De Vos M, Boylan G, Ward T. Multivariate generative adversarial networks and their loss functions for synthesis of multichannel ECGs. IEEE Access. 2021;9:158936-45. https://doi.org/10.1109/ACCESS.2021.3129471.

Lou J, Wang Y, Li L, Zeng D. Learning latent heterogeneity for type 2 diabetes patients using longitudinal health markers in electronic health records. Stat Med. 2021;40(8):1930-46. https://doi.org/10.1002/sim.8887.

Gopalan A, Mishra P, Alexeeff SE, Blatchins MA, Kim E, Man AH, et al. Prevalence and predictors of delayed clinical diagnosis of type 2 diabetes: a longitudinal cohort study. Diabet Med. 2018;35(12):1655-62. https://doi.org/10.1111/dme.13712.

Urina-Jassir M, Herrera-Parra LJ, Hernandez Vargas JA, Valbuena-García AM, Acuña-Merchán L, Urina-Triana M. The effect of comorbidities on glycemic control among Colombian adults with diabetes mellitus: a longitudinal approach with real-world data. BMC Endocr Disord. 2021;21(1):128. https://doi.org/10.1186/s12902-021-00789-y.

Bing S, Dittadi A, Bauer S, Schwab P. Conditional generation of medical time series for extrapolation to underrepresented populations. PLOS Digit Health. 2022;1(7):e0000074. https://doi.org/10.1371/journal.pdig.0000074.

Brophy E, Wang Z, She Q, Ward T. Generative adversarial networks in time series: A survey and taxonomy. arXiv [Preprint]. 2021:arXiv:2107.11098. https://doi.org/10.48550/arXiv.2107.11098.

Liu D, Wu Y, Hong D, Wang S. Time series data augmentation method of small sample based on optimized generative adversarial network. Concurr Comput Pract Exp. 2022;34(27):e7331. https://doi.org/10.1002/cpe.7331.

Larrea X, Hernandez M, Epelde G, Beristain A, Molina C, Alberdi A, et al. Synthetic subject generation with coupled coherent time series data. Eng Proc. 2022;18(1):7. https://doi.org/10.3390/engproc2022018007.