UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS
Curso de Graduação em Farmácia-Bioquímica

DADOS SINTÉTICOS EM SAÚDE

Da privacidade de dados do paciente ao futuro da pesquisa com dados longitudinais

VITOR GALVÃO LOPES

Trabalho de Conclusão do Curso de Farmácia-Bioquímica da Faculdade de Ciências Farmacêuticas da Universidade de São Paulo.

Orientador(a):
Prof.(a). Dr(a) Gabriel L. B. de Araujo

SÃO PAULO – SP

2024

Sumário

LISTA DE ABREVIATURAS

RWD	*Real World Data* ou dados de mundo real
TIC	Tecnologias de Informação e Comunicação
LGPD	Lei Geral de Proteção de Dados
IoT	*Internet-of-things* ou Internet das coisas
TSTR	Treinar no Sintético, Testar no Real
TRTS	Treinar no Real, Testar no sintético
EHRs	Electronic Health Records ou Registros Eletrônicos de Saúde
PHI	Protected Health Information, ou Informação de Saúde Protegida
TI	Tecnologia da Informação
GAN	Redes Adversariais Generativas
SMOTE	Técnica de Sobre Amostragem Minoritária Sintética
VAEs	Variational Autoencoders ou Autocodificadores Variacionais
PCA	Análise de Componente Principal
CGAN	Rede Adversarial Gerativa Condicional
WGAN	Rede Adversarial Generativa de Wasserstein
WGANGP	WGAN com penalidade de gradiente
AC-GAN	Rede Adversarial Gerativa de Classificadores Auxiliares
OSIM	Observational Medical Outcomes Partnership
OMOP	Common Data Model
CDM	Simulator ou Simulador de Conjunto de Dados Médicos Observacionais
IMS	Correspondências Exatas entre Dados Sintéticos e Originais
NNDR	Taxa de Distância do Vizinho Mais Próximo
CSF	Fidelidade Sintética Clínica
GSF	Fidelidade Sintética Genômica
SHAP	Shapley Additive Explanations
CART	Classification and Regression Trees
MedGaN	Medical Generative Adversarial Network
XGBoost	Extreme Gradient Boosting
MtGaN	Medical Text Generative Adversarial Network ou Rede Adversarial Geradora de Texto Médico
DRAI	Dual Adversarial Inference ou Inferência Adversarial Dupla
CaVAe	Conditional Adversarial Variational Autoencoder
DAI	Dual Adversarial Inference ou Inferência Adversarial Dupla
BTF	Bayesian Tensor Factorization ou Fatoração por Tensor Bayesiano
MIA	Membership Inference Attack ou Ataque de Inferência de Associação
DLA	Data Labelling Analysis ou Análise de Rotulação de Dados
MAE	Mean Absolute Error ou Erro Médio Absoluto
DAG	Directed Acyclic Graphs ou Gráficos Acíclicos Direcionados
ML-PK	Machine Learning - Pharmacokinetics
LSTM	Long Short-Term Memory ou Memória de Curto-Longo Prazo
DP-GAN	Differentially Private Generative Adversarial Network

RTSGAN	Real-World Time Series Generative Adversarial Network
t-SNE	Stochastic Neighbor Embedding
DCGAN	Deep Convolutional Generative Adversarial Network
WGAN-GP-SN	Wasserstein GAN with Gradient Penalty and Spectral Normalization
DSC	Dice Similarity Coefficient ou Coeficiente de Similaridade de Dados
HD th	Percentile Hausdorff Distance ou Distância de Hausdorff do º Percentil
FID	Fréchet Inception Distance ou Distância Inicial de Frechet
MSPNs	Mixed Sum-Product Networks ou Redes de Soma Mista de Produtos
PoAC	Proportion of Alternatives Considered
GK-MMD	Graph Kernel Maximum Mean Discrepancy
LOSO	Leave One Subject Out Cross-Validation
LR	Logistic Regression Model ou Modelo de Regressão Logística
dBms	Deep Boltzmann Machines ou Máquinas Boltzmann Profundas
MICE	Multivariate Imputation by Chained Equations
GPT	Generative Pre-trained Transformer
FL	Federated Learning ou Métodos de Aprendizagem Federada
VAMBN	Variational Autoencoder Modular Bayesian Network of the VAMBN
MST	Minimum Spanning Tree ou Método MST
HME	Homomorphic Encryption ou Criptografia Homomórfica
pGAN	Private Generative Adversarial Networks ou Modelo GAN Privativo
BilsTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network ou Rede Neural Convolucional
MPoM	Mixture of Product of Multinomials
CLGP	Categorical Latent Gaussian Process
EMR	Electronic Medical Records ou Registros Médicos Eletrônicos
CEP	Código de Endereçamento Postal
GDPR	General Data Protection Regulation
DP	Differential Privacy ou Privacidade Diferencial

RESUMO

LOPES, VG. Dados Sintéticos em Saúde: Da privacidade de dados do paciente ao futuro da pesquisa com dados longitudinais. 2024. Trabalho de Conclusão de Curso de Farmácia-Bioquímica – Faculdade de Ciências Farmacêuticas – Universidade de São Paulo, São Paulo, 2024.

Palavras-chave: Dados Sintéticos; Privacidade; Dados de mundo real; Dados médicos; Inteligência Artificial.

INTRODUÇÃO: Com o avanço de novos algoritmos de predição e de inteligência artificial, a disponibilidade de dados de mundo real (RWD) se torna um fator limitante para o desenvolvimento de pesquisas em saúde. Apesar da grande geração de dados atual há limitação no acesso devido à natureza sensível dos dados. **OBJETIVO:** O presente trabalho tem como objetivo analisar o impacto do uso de dados sintéticos para a realização de investigações científicas e regulatórias na área médica, garantindo a segurança da informação e a reprodutibilidade dos estudos. Além disso, busca discutir a necessidade de padronização de métodos para validação de resultados e a adaptação das estruturas regulatórias ao uso de dados sintéticos. **MATERIAIS E MÉTODOS:** A metodologia adotada consiste em uma revisão bibliográfica sobre o uso de dados sintéticos na área da saúde, com foco em aplicações reais e discussão dos limites de privacidade. A coleta de dados foi realizada nas bases PubMed e Scopus, utilizando palavras-chave relacionadas ao tema. Foram analisados estudos que tratam da geração de dados sintéticos, sua aplicação em pesquisas e a discussão sobre desafios regulatórios e de privacidade. Os métodos de análise incluíram a avaliação crítica da literatura disponível e a interpretação dos resultados quanto à aplicabilidade e segurança desses dados no campo da saúde. **RESULTADOS:** A análise dos trabalhos demonstra que os dados sintéticos têm potencial para acelerar o desenvolvimento de estudos com dados de mundo real (RWD), contribuindo para avanços na saúde. No entanto, os métodos de avaliação de qualidade e segurança demonstram variações entre os estudos, o que dificulta a comparação entre propostas. **CONCLUSÃO:** Embora os dados sintéticos sejam uma ferramenta promissora para a área da saúde, ainda há desafios a serem superados, como a padronização dos métodos de validação de performance e privacidade.

Somado a isso, a ausência de diretrizes regulatórias claras implica em obstáculos para adoção dos dados sintéticos em larga escala, sendo necessário esforços contínuos para estabelecer critérios de privacidade e segurança dos métodos de validação e garantir a plena utilidade e manutenção da privacidade dos pacientes.

ABSTRACT

Keywords: Synthetic Data; Privacy; Real-World Data; Medical Data; Artificial Intelligence.

INTRODUCTION: With the advancement of new prediction algorithms and artificial intelligence, the availability of real-world data (RWD) has become a limiting factor for the development of health research. Despite the large amount of data being generated today, access remains restricted due to the sensitive nature of health data. OBJECTIVE: This study aims to analyze the impact of using synthetic data for scientific and regulatory investigations in the medical field, ensuring information security and the reproducibility of studies. Additionally, it seeks to discuss the need for standardizing methods for result validation and the adaptation of regulatory frameworks to the use of synthetic data. MATERIALS AND METHODS: The adopted methodology consists of a literature review on the use of synthetic data in healthcare, focusing on real-world applications and the discussion of privacy limits. Data collection was conducted in the PubMed and Scopus databases, using keywords related to the topic. Studies addressing the generation of synthetic data, its application in research, and the discussion of regulatory and privacy challenges were analyzed. The methods of analysis included a critical assessment of the available literature and the interpretation of the results regarding the applicability and security of such data in the health field. RESULTS: The analysis of the studies shows that synthetic data has the potential to accelerate the development of studies with real-world data (RWD), contributing to advances in healthcare. However, quality and safety assessment methods vary across studies. CONCLUSION: Although synthetic data is a promising tool for the healthcare sector, there are still challenges to be overcome, such as the establishment and standardization of performance and privacy validation methods. In addition, the absence of clear regulatory guidelines presents obstacles to the large-scale adoption

of synthetic data, requiring continuous efforts to establish privacy and security criteria for validation methods, ensuring full utility and the preservation of patient privacy.

1. INTRODUÇÃO

1.1. Informação e privacidade de dados

O ritmo atual de acumulação de dados juntamente ao desenvolvimento de algoritmos avançados para obtenção de informação tem revolucionado os métodos de tomada de decisão, pesquisa e inovação em diversas áreas do conhecimento.

Em especial, a integração de informações pelo fenômeno de Tecnologias de Informação e Comunicação (TIC) oferece novas possibilidades para melhorar a jornada do paciente e reduzir custos operacionais (MAMLIN; TIERNEY, 2016). As principais áreas de desenvolvimento das TIC incluem registos de saúde eletrônicos, troca de informações por meio de dispositivos de *Internet-of-things (IoT)*, portais de pacientes, telemedicina e dispositivos *wearable*, como smartphones e *smartwatches* (ACETO; PERSICO; PESCAPÉ, 2018; MAMLIN; TIERNEY, 2016). Estas tecnologias permitem a monitorização remota dos pacientes, visitas virtuais e um maior envolvimento dos pacientes através de aplicações móveis de saúde (DOWNES; HORIGAN; TEIXEIRA, 2019).

No entanto, apesar da geração de um grande volume de dados, a transformação de informação em tomada de decisão depende da disponibilidade dos dados para treino e execução de algoritmos especializados. Tratando-se do escopo da saúde, frequentemente são os dados são de origem confidenciais, contendo informações sensíveis sobre o estado de saúde do paciente O armazenamento e manipulação destes dados levanta preocupações sobre a violação de direitos fundamentais de privacidade, resguardados por legislações regulatórias contemporâneas como a Lei Geral de Proteção de Dados (LGPD) (FERREIRA et al., 2022) no Brasil e o *General Data Protection Regulation (GDPR)* na União Europeia (FATEHI et al., 2020; FIGUEIREDO; VARELLA, 2022).

1.2. Marco Regulatório de proteção de dados

A LGPD no Brasil foi um marco jurídico significativo, no entanto, as organizações enfrentam desafios para cumprir integralmente a lei. O estudo de (FERREIRA et al., 2022) constatou que muitas empresas estão iniciando projetos de adaptação à LGPD, mas as estruturas internas permanecem pequenas e exigem treinamento adicional dos funcionários. A LGPD, juntamente com a Emenda Constitucional nº 115/2022, introduziu sistemas de regulação e governança para proteção de dados nos setores público e privado, havendo necessidade de maior densidade constitucional e parâmetros definidos de aplicabilidade na ordem jurídica constitucional para controlar a produção de atos normativos, à medida que o Brasil adota o constitucionalismo digital para salvaguardar os direitos humanos fundamentais contemporâneos (PAIVA; LANZILLO, 2024). Por sua vez, a saúde é considerada um dos setores mais complexos para a conformação com a regulação devido ao constante tratamento de dados sensíveis (HAWRYLISZYN; COELHO; BARJA, 2021). A implementação destas leis impacta a equipes de desenvolvimento de software, particularmente na elicitação de requisitos de privacidade (CANEDO et al., 2022). A implementação adequada requer envolvimento institucional, boas práticas, normas de segurança, padrões técnicos, ações educativas, auditorias internas e estratégias de mitigação de riscos (HAWRYLISZYN; COELHO; BARJA, 2021). Tópicos de pesquisa emergentes neste campo incluem aplicação de blockchain e e de aprendizado de máquina para mitigar as questões de segurança (FATEHI et al., 2020).

As técnicas atuais para promover a privacidade dos dados dos pacientes nos cuidados de saúde incluem a encriptação, geração de chaves multibiométricas, encriptação do tipo *attribute-based* e a anonimização/pseudonimização de dados (MASOOD et al., 2018; SAHI et al., 2017). Em contraste as redes centralizadas e físicas, a introdução de computação em nuvem impôs novos desafios, necessitando de medidas de segurança ao longo de todo o ciclo de vida dos dados de saúde, desde a recolha de dados, desde a coleta, o armazenamento e o acesso (BOSE; MARIJAN, [s.d.]). As ferramentas de anonimização existentes podem ser suscetíveis a ataques, e a gestão dessas informações sensíveis representa um desafio constante para na busca por aceleração do conhecimento científico ou melhoramento de serviços, o que

levou pesquisadores a explorarem novas formas de trabalharem com dados sensíveis por meio de dados sintéticos (GIUFFRÈ; SHUNG, 2023).

1.3. Definição de dado sintético

Sendo uma área do conhecimento em expansão, há divergências sobre a definição de dados sintéticos, uma definição abrangente segundo o *The Royal Society and The Alan Turing Institute*, em tradução livre “dados que foram gerados usando um modelo matemático ou algoritmo construído com o propósito de resolver uma (ou um conjunto de) tarefa(s) de ciência de dados” (JORDON et al., 2022).

Os dados sintéticos são gerados artificialmente para imitar dados do mundo real e têm demonstrado utilidade para rotinas de inteligência artificial (IA) e na garantia de privacidade de dados uma vez que não contém informações pessoais. Nesse escopo, há o intuito de aumentar a disponibilidade de dados para pesquisa científica e reduzir vieses nos modelos de aprendizagem de máquina (MARWALA; FOURNIER-TOMBS; STINCKWICH, 2023). Os dados sintéticos podem ser usados para treinar modelos de IA, pesquisa clínica e educação médica, promovendo a disponibilidade de acesso aos dados e ao mesmo tempo que protege a privacidade do paciente (ARORA; ARORA, 2022). No entanto, existem desafios para garantir a qualidade e autenticidade dos dados que são submetidos a testes para avaliar se o dado sintético pode representar fielmente os dados originais e garantir privacidade (VALLEVIK et al., 2024). Por outro lado, avaliações empíricas mostraram que os modelos treinados com dados sintéticos podem ter um desempenho comparável àqueles treinados em dados reais para diversas tarefas de aprendizado de máquina (HITTMEIR; EKELHART; MAYER, 2019).

1.4. Características fundamentais de dados sintéticos

Por meio das definições trazidas pelo Instituto *Alan Turing* (JORDON et al., 2022), a geração destes dados necessita de um gerador de dados sintéticos (SDG) que deve conter propriedades específicas para satisfazer requisitos fundamentais. De acordo com (JORDON et al., 2022), o gerador sintético de dados deve apresentar características de:

- Precisão sintática: Os dados gerados devem ser plausíveis, ou seja, o dado deve respeitar certas propriedades estruturais, como por exemplo

um Código de Endereçamento Postal (CEP) gerado sinteticamente deve poder existir, dada as regras de criação de um CEP. Outro caso envolve geração de dados de séries temporais, é necessário garantir que os estes não sejam gerados usando informações do futuro.

- Privacidade: Deve ser possível quantificar precisamente quanta informação sobre os dados originais é revelada através da liberação da amostra sintética, e o método empregado para medição de privacidade dependerá da tarefa específica que o dado será empregado. Embora a abordagem de privacidade diferencial seja uma maneira popular de avaliar a quantidade de informação liberada através de geradores de dados sintéticos, uma noção diferente pode ser necessária quando os dados são esparsos ou se deseja se afastar de limites de pior caso.
- Precisão Estatística: Deve ser possível quantificar precisamente a semelhança estatística entre os dados sintéticos e os originais. Um bom gerador de dados sintéticos deve permitir controle sobre certas distribuições marginais e certas relações entre variáveis.
- Eficiência: O algoritmo deve ser capaz de escalar à medida que há aumento de dimensão do espaço de dados (dimensionalidade). A consideração do uso de dados sintéticos pode indicar que os dados originais são inadequados para a tarefa em questão, seja porque são privados, enviesados ou diminutos. Portanto, dados sintéticos que sejam muito semelhantes aos dados originais também sofrerão dos mesmos problemas, sendo assim o grau de similaridade permitido constitui-se em três atributos fundamentais para a geração de dados sintéticos: utilidade, fidelidade e privacidade. As próximas definições foram parafraseadas do trabalho sobre dados sintéticos do Instituto Alan Turing (JORDON et al., 2022).
- Utilidade: A utilidade dos dados sintéticos geralmente é determinada por sua capacidade de satisfazer uma tarefa ou conjunto de tarefas específicas. Isso frequentemente envolve comparar o desempenho de modelos treinados com dados reais versus dados sintéticos e pode envolver a inspeção de métricas concretas como exatidão (*accuracy*), precisão (*precision*), entre outros. Fazer isso frequentemente requer o paradigma Treinar no Sintético, Testar no Real (TSTR) e Treinar no Real, Testar no sintético (TRTS) (ESTEBAN; HYLAND; RÄTSCH,

2017), no qual os modelos são treinados com dados sintéticos e seu desempenho é avaliado com dados reais.

- Fidelidade: A definição de fidelidade como medidas que comparam diretamente o conjunto de dados sintéticos com o real (em vez de indiretamente através de um modelo ou desempenho em uma tarefa específica). De uma perspectiva de alto nível, a fidelidade é o quão bem os dados sintéticos correspondem estatisticamente aos dados reais. No caso mais geral, a similaridade estatística completa (ou seja, correspondência das distribuições dos dados sintéticos e reais) deve permitir que muitas tarefas que seriam realizadas com os dados reais sejam realizadas com os dados sintéticos. No entanto, tal correspondência é difícil, especialmente na presença de requisitos de privacidade (ULLMAN; VADHAN, 2011), e até indesejável na presença de vieses (VAN BREUGEL et al., 2021).

Há uma relação inversa entre fidelidade e privacidade, mas não necessariamente entre utilidade e privacidade. Dessa forma é possível manipular a geração de dados para situações específicas para a resolução de problemas específicos.

1.5. Aplicações de dados sintéticos

O uso de dados sintéticos no contexto de saúde está focado em criar dados novos a partir de características estatísticas dos dados originais, geralmente com o objetivo de substituir dados reais e permitir acesso facilitado para pesquisa e uso em modelos de ciência de dados (GIUFFRÈ; SHUNG, 2023). Não somente, a utilização de dados sintéticos é empregada para melhorar a inferência em tarefas de inteligência artificial, na disciplina de visão computacional e imagens médicas, os dados sintéticos podem melhorar os modelos de *deep learning* (aprendizado profundo), por meio do aumento da diversidade de dados e redução do desequilíbrio entre classes (PAPROKI; SALVADO; FOOKES, 2024).

Os principais métodos incluem técnicas clássicas de desidentificação e modelos baseados em aprendizagem profunda, particularmente Redes Adversariais Generativas (GANs) (NIK et al., 2023). Outros métodos importantes para a criação de dados sintéticos incluem técnicas baseadas em Análise de componente principal (PCA), técnica de sobre amostragem minoritária sintética (SMOTE), *variational*

autoencoders (VAEs), com SMOTE e PCA apresentando melhor desempenho na reprodução de características de dados originais (ALLEN; SALMON, 2020). As GANs demonstraram um desempenho notável na geração de conjuntos de dados tabulares complexos, preservando ao mesmo tempo as características estatísticas e a privacidade (NIK et al., 2023). Os dados sintéticos têm diversas aplicações na área da saúde, incluindo pesquisa de simulação, teste de algoritmos, epidemiologia, desenvolvimento de tecnologia da informação (TI) em saúde, educação e vinculação de dados (Gonzales et al., 2023). No entanto, a avaliação de métodos de geração de dados sintéticos continua a ser um desafio devido à falta de um quadro universal de benchmarking (Yan et al., 2022; Hernandez et al., 2022).

Nesse contexto, os dados sintéticos podem viabilizar pesquisa utilizando dados médicos, inclusive dos nomeados dados de mundo real (*Real World Data*), que são dados sobre o estado de saúde dos pacientes ou prestação de cuidados de saúde coletados rotineiramente por diversas fontes (FOOD AND DRUG ADMINISTRATION, 2023). Estas informações compreendem dados de dispositivos móveis de monitorização, contas médicas, *eletronic health records* (EHRs), entre outros. Em vista disso, a geração de dados sintéticos viabilizou casos de avaliação de políticas públicas de saúde (HENNESSY, 2015), avaliação tratamentos e intervenções clínicas em saúde (ENANORIA et al., 2016) e permite avanços em medicina personalizada pela modelagem e refinamento de modelos com populações específicas (NGUFOR et al., 2019).

2. MATERIAIS E MÉTODOS

Para conduzir uma revisão abrangente do estado da arte da tecnologia de geração de dados sintéticos, com foco nas aplicações na área da saúde foram definidos os seguintes objetivos:

Q1: Qual é o panorama de aplicações de dados sintéticos em saúde? Considerando foco principalmente em dados longitudinais.

Q2: Quais os principais desafios e ganhos em termos de segurança da informação de pacientes com dados sintéticos?

Q3: Quais são as perspectivas futuras para a tecnologia e como isso impacta a pesquisa com dados de mundo real?

O levantamento de referências bibliográficas utilizou a busca automática por termos nas bases de dados de publicações Scopus e PubMed. Os termos foram organizados em três pilares principais de busca: *Synthetic Data*, *Patient Privacy* e *Medical Data*. Essa etapa foi sucedida por uma busca manual utilizando critérios de inclusão e exclusão ao avaliar os materiais pelo título e abstract. Foram considerados apenas artigos científicos revisado em pares (*peer-review*), que abordam modelos e sistemas utilizando dados médicos longitudinais, escritos na língua inglesa ou portuguesa e publicados a partir do ano de 2015. Os critérios de exclusão desconsideraram artigos dedicados ao estudo de dados sintéticos para outras espécies que não a humana, estudos nos quais o texto completo não foi acessível e artigos de conferência. Após a aplicação dos critérios de seleção foram mantidos 53 artigos que contemplaram as características propostas (Figura 1).

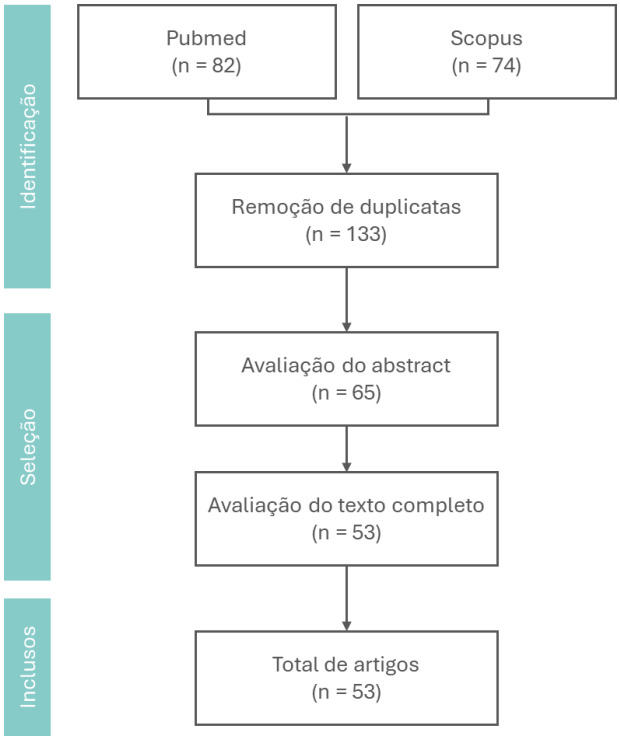

Figura 1 – Fluxo de seleção dos artigos para a revisão

3. RESULTADOS

A possibilidade de geração de dados sintéticos tem impactado a área da saúde por facilitar o acesso a dados que antes eram inacessíveis. Os artigos analisados destacam avanços significativos nas arquiteturas para a geração de dados sintéticos, contudo, a área ainda enfrenta desafios relacionados à padronização dos métodos de validação de performance e inviolabilidade de informações de saúde do paciente (PHI), o que revela uma coleção científica de caráter primariamente exploratório.

3.1. Critérios de avaliação de performance em dados sintéticos

A falta de métodos sistemáticos para avaliar o realismo e a validade de dados sintéticos de registros de saúde foi abordada entre os trabalhos selecionados (KIM et al., 2024; LANGE; N; E, 2024; LITTLE; ELLIOT; ALLMENDINGER, 2023; WALONOSKI et al., 2018). Para tal foi construída uma tabela (Tabela 1) com os métodos de geração de dados sintéticos e os principais métodos de avaliação de desempenho dos modelos. Os grandes grupos de métricas observados, entre outros, foram:

- Métricas de aprendizado de máquina: exatidão (*accuracy*), precisão (*precision*), *score* F1, recall, curva AUC, matriz de confusão
- Similaridade estatística: Teste U de Mann-Whitney, Teste de Kolmogorov-Smirnov, Índices de similaridade de Jaccard, Teste qui-quadrado de Pearson, Teste de soma de classificação de Mann-Whitney, Estimativa de máxima verossimilhança (MLE), Coeficiente de similaridade de dados (DSC), Distância de Hausdorff do 95º percentil (95HD), Distância inicial de Frechet (FID)
- Métricas de privacidade: Salvaguardas de privacidade diferencial (DP), Validação cruzada de Leave One Subject Out (LOSO), Distância de registros mais próximos, Teste de inferência de associação, *Singling Out*; Anonímetro, Vinculabilidade; Inferência de atributos, PoAC (Proporção de alternativas consideradas)
- Séries temporais: Erro quadrático médio (RMSE)

Tabela 1 – Principais métodos de geração de dados abordados nos estudos e as técnicas de validação de desempenho (D) e privacidade (P).

Principais métodos de geração de dados	**Principais técnicas de validação citadas**	**Artigo**
- Redes adversárias generativas (GANs) - Autocodificadores variacionais (VAEs) - Árvores sequenciais - Tecnologias de aprimoramento da privacidade (criptografia homomórfica, blockchain) - Aprendizagem federada	-	(ALLOZA et al., 2023)
- Rede Adversarial Generativa (GAN) - Rede Adversarial Gerativa Condicional (CGAN) - Rede adversária generativa de Wasserstein (WGAN) com penalidade de gradiente (WGANGP)	- Análise de perdas, Comparação de correlações, Medição de similaridade, Teste U de Mann-Whitney, Teste de Kolmogorov-Smirnov (K-S) e Índices de similaridade de Jaccard (desempenho)	(ARVANITIS et al., 2022)
- Redes adversárias generativas (GANs) - Rede Adversarial Gerativa de Classificadores Auxiliares (AC-GAN) - Privacidade diferencial	- Comparação de distribuições variáveis estrutura de correlação entre dados reais e simulados, Avaliação cega de dados de nível individual por médicos, Avaliação das decisões de tratamento com base em participantes sintéticos, Medição da correlação entre os parâmetros da visita de estudo (D)	(BEAULIEU-JONES et al., 2019)
- Sistema MDClone - Sistema Synthea - Simulador de conjunto de dados médicos observacionais (OSIM) - Codificadores automáticos - Redes adversárias geradoras (MedGaN)	- Comparação sistemática de 5 estudos observacionais; Comparação dos resultados obtidos de dados sintéticos com os de dados originais; Análise estatística para estimar viés e estabilidade; Avaliação da consistência das estimativas em conjuntos sintéticos, uso de modelos multivariados para avaliar resultados (D)	(BENAIM et al., 2020)
- Synthea (software)	-	(CHEN et al., 2019; DIOUF et al., 2024)

- Arquitetura de rede adversária generativa condicional (GAN) - GAN tabular condicional de Wasserstein	- Correspondências exatas entre dados sintéticos e originais (IMS) (P) - Taxa de distância do vizinho mais próximo (NNDR), fidelidade sintética clínica (CSF), Fidelidade sintética genômica (GSF), Explicações sobre aditivos de Shapley (SHAP), Curvas de Kaplan-Meier para análise de sobrevivência, Risco proporcional de Cox e modelos de regressão de Cox penalizados por L1 (D)	(D'AMICO et al., 2023)
- Síntese sequencial usando árvores de decisão - Rede adversária generativa condicional - Codificador automático variacional	- Métrica de divulgação de atribuição, métrica de divulgação de associação (P) - Estimar concordância, Acordo de decisão, Sobreposição de intervalo de confiança (D)	(EL KABABJI et al., 2023)
- Simulação estatística - Derivação computacional	- Teste qui-quadrado de Pearson,Teste de soma de classificação de Mann-Whitney, Testes de soma de classificação de Wilcoxon, Coeficientes de correlação de classificação de Spearman, Validação cruzada de 5 vezes (D)	(FORAKER et al., 2020)
- Modelos de árvores de classificação e regressão (CART) - Redes adversárias generativas (GANs)	- Evidência estatística para validar atributos relevantes - Classificadores baseados em aprendizado de máquina para provar desempenho equivalente (D)	(GALLOS et al., 2024)
- GAN, MedGan - Imputação múltipla - Imputação multivariada por equações em cadeia (MICE)	- Divulgação de associação, Divulgação de atributos (P) - Métricas de classificação cruzada (D)	(GONCALVES et al., 2020)
- Redes adversárias generativas (GANs)	- Validação por oncologistas usando a taxa de mortalidade indireta obtida para pacientes em diferentes estágios (D)	(GONZALEZ-ABRIL et al., 2021)
- MDClone	- Regressão logística, Máquinas de aumento de gradiente extremo (XGBoost), Testes de qui-quadrado, testes t, Testes U (Mann-Whitney) (D)	(GREENBERG et al., 2023)

- Rede Adversarial Geradora de Texto Médico (MtGaN), GAN condicional - Algoritmo REINFORCE - Rede Adversarial Generativa (GAN) - Procura de recompensas intermediárias em Monte Carlo	- Avaliação adversarial, Erro de confiabilidade do avaliador (ERE), Estimativa de máxima verossimilhança (MLE) (D)	(GUAN et al., 2021)
- Redes geradoras adversárias condicionais (CGANs) - Inferência adversarial dupla (DRAI) - InfoGan - Autocodificador Variacional Adversarial Condicional (CaVAe) - Inferência adversarial dupla (DAI) - Modelo baseado em U-Net	- Avaliação de geração condicional, Testes de recuperação de imagem, Medição de erro de desemaranhamento, Erro de consistência de características de imagem cruzada (CIFC), Divergência de discordância, sobreposição de rótulos (D)	(HAVAEI et al., 2021)
- Rede Adversarial Gerativa Condicional (CGaN) - Fatoração por tensor bayesiano (BTF)	- Curva de precisão e recuperação (PR), Área sob a curva (AUC) para curvas ROC e PR (D)	(HUANG et al., 2024)
-Wasserstein GAN com penalidade de gradiente (WGAN-GP) -DoppelGANger	- Ataque de inferência de associação (MIA) (privacidade) - AUC-ROC, Análise de rotulagem de dados (DLA), Erro médio absoluto (MAE), Métrica de similaridade de cosseno, testes t para avaliação de significância (D)	(ISASA et al., 2024)
- Método bayesiano de geração de dados	- Modelo bayesiano para distribuições posteriores, Conjunto profundo para estimativa da incerteza do modelo, Técnicas de inferência causal usando gráficos acíclicos direcionados (DAG), Modelo híbrido de farmacocinética populacional (ML-PK) de aprendizado de máquina (D)	(JANSSEN et al., 2024)
- Redes adversárias generativas (GANs) - CT-GAN - GAN com implementação de memória de curto-longo prazo (LSTM-GAN) - DP-GAN	- TSTR, TRTS (D) - Inspeção visual (VI) e Protocolo de marcação d'água de dados (DWP) (P)	(KHAN; MURTAZA; AHMED, 2024)

- Redes adversárias geradoras de séries temporais do mundo real (RTSGAN)	- Distância de registros mais próximos, Teste de inferência de associação (P) - AUC; Erro quadrático médio de propensão (MSE), análise de incorporação estocástica de vizinhos (t-SNE), Análise do histograma (D)	(KIM et al., 2024)
- GAN convolucional profundo (DCGAN) - Wasserstein GAN com penalidade de gradiente (WGAN-GP) - Wasserstein GAN com normalização espectral (WGAN-GP-SN)	- Coeficiente de similaridade de dados (DSC), Distância de Hausdorff do 95º percentil (95HD), Distância inicial de Frechet (FID) (D)	(KOSSEN et al., 2021)
- Redes de soma mista de produtos (MSPNs)	- Proporção de alternativas consideradas (PoAc) (P) - Análise de regressão, Comparação da distribuição da amostragem empírica (D)	(KROES et al., 2022)
- CART (Árvores de inferência condicional) - Métodos bayesianos não paramétricos - Técnicas de rede neural (e.g máquinas Boltzmann restritas empilhadas)	- Métrica de proximidade t (privacidade)	(KUIPER; VAN DEN HEUVEL; SWERTZ, 2015)
- Redes adversárias generativas (GANs) - Proteções de privacidade diferencial (DP)	- Salvaguardas de privacidade diferencial (DP) (P) - Validação cruzada de Leave One Subject Out (LOSO), Modelo de regressão logística (LR) (D)	(LANGE; N; E, 2024)
- Redes adversárias generativas (GANs)	- Modelo codificador-decodificador, Medidas de sobreposição de N gramas, Sensibilidade e valor preditivo positivo (PPV), Avaliação da validade epidemiológica (D)	(LEE, 2018)
- Máquinas Boltzmann profundas (dBms) - Autocodificadores variacionais (VAEs) - Redes adversárias generativas (GANs) - Imputação multivariada por equações em cadeia (MICE) - Marginais independentes (IM)	- Métricas de privacidade de membros, Medidas de privacidade diferenciais (P) - Análise de sobreajuste, RMSE de razões de chances logarítmicas (D)	(LENZ; HESS; BINDER, 2021)

- memória de curto-longo prazo (LSTM) - GPT-2	- Estudo de usuário para avaliação de privacidade, Recordação ROUGE-N para medição de similaridade, Pontuação BM25 para recuperação de documentos (P) - Precisão em nível de entidade, recall e pontuação F1 para medição de desempenho, Comparação de modelos treinados em dados reais, mistos e sintéticos (D)	(LIBBI et al., 2021)
- Métodos de aprendizagem federada (FL) - Redes adversárias generativas (GAN) - Autocodificadores variacionais (VAE) - GANs condicionais - Modelo de mistura gaussiana bayesiana variacional federada	-	(LITTLE; ELLIOT; ALLMENDINGER, 2023)
- Duplicação de conjuntos de dados - Inserção de erros (inserção, exclusão, substituição) no atributo do sobrenome - Remoção de diferentes atributos dos conjuntos de dados	-	(MAMUN; ASELTINE; RAJASEKARAN, 2016)
- Synthea (software)	- Testes de qui-quadrado para variáveis binárias,Testes t para variáveis contínuas, Comparação com os resultados de estudos publicados, Rastreamento de problemas e implementação de soluções (D)	(MEEKER et al., 2022)
- VAMBN - MultiNODEs	- Jensen-Shannon-Divergence (JSD), Preservação da estrutura de correlação entre variáveis nos dados sintéticos (D) - Pontuações de risco de privacidade avaliando risco de diferenciação, risco de vinculabilidade e risco de inferência (P)	(MOAZEMI et al., 2024)
- Synthea (Software)	- Acurácia de predição (D)	(MOHAMAD et al., 2022)

- Autocodificadores variacionais (VAEs) - Redes adversárias generativas (GANs)	- Medição da precisão da rede de reconhecimento de identidade (P) - Medição da precisão da rede de reconhecimento de doenças - SSIM (Índice de Similaridade Estrutural) para avaliação de realismo, Score F1 para reconhecimento de doenças (D)	(MONTENEGRO; CARDOSO, 2024)
- Synthea - Redes adversárias generativas (GANs) - Bayes variacionais de codificação automática (AEVB)	- Privacidade diferencial, Geração de dados sintéticos, Aprendizagem federada (P) - Abordagem de gerenciamento baseada em risco (D)	(NAIK et al., 2024)
- Codificadores automáticos de gráficos variacionais (VGAEs)	- Máquina de vetores de suporte (SVM) de uma classe para identificação de valores discrepantes, métrica de discrepância média máxima do kernel do gráfico (GK-MMD) para avaliar a qualidade do gráfico, Ajuste de variância da distribuição gaussiana para garantir a novidade das amostras geradas (D)	(NIKOLENTZOS et al., 2023)
- Métodos de análise de dados espaço-temporais - Modelo bayesiano	-Avaliação de risco com base nos riscos de divulgação esperados (P) - Algoritmo MCMC para ajuste de modelos hierárquicos, Comparação de disparidades urbanas/rurais (D)	(QUICK; WALLER, 2018)
- Software estatístico R com o pacote synthpop	- Análise ROC, Teste t de 2 amostras, Teste Ryan-Joiner para normalidade (D)	(RASHIDI et al., 2022)
- Modelagem de ocorrência e data da morte - Análise da árvore de classificação - Modelo de risco proporcional de Cox	- Comparação das estimativas dos parâmetros de sobrevivência entre dados sintéticos e dados reais, Avaliação das distribuições e concordância percentual de mortalidade por todas as causas e mortalidade por causas específicas; estatística Kappa para medir a concordância de parâmetros estatisticamente significativos (D)	(RESNICK; COX; MIREL, 2021)

- Autocodificadores variacionais (VAEs) - Redes adversárias generativas (GANs) - Modelos probabilísticos de difusão de redução de ruído (DDPMs)	- Classificador CatBoost (D) - Distância até o vizinho mais próximo usando a distância de Hamming (P)	(RÖCHNER, 2024)
- Pacote de software Faker - Suíte de ferramentas Posda - Pacote de software ImageMagick	- Comparação com uma chave de resposta para validar os resultados de desidentificação, Verificações automatizadas de SOPs duplicados e inconsistências (D) - Revisão visual de imagens para possíveis PHI, Revisão das tags DICOM para remoção de PHI (P)	(RUTHERFORD et al., 2021)
- Modelos de difusão	- Distância inicial de Fréchet (FID), Similaridade estrutural em várias escalas (MS-SSIM), Pontuação de dados (DS) e Interseção sobre Union (IoU) (D)	(SARAGIH; HIBI; TYRRELL, 2024)
- VAMBN (Rede Bayesiana Modular de Autoencoder Variacional) - VAMBN-MT (extensão de pontos de tempo memorizados do VAMBN)	- VAMBN-MT (D) - Singling Out; Anonímetro; Vinculabilidade; Inferência de atributos (P)	(SCHNEIDER et al., 2024)
- Emulação da estrutura de covariância e curva de sobrevivência	- Comparação de estimativas de sobrevivência por todas as causas entre dados reais e dados sintéticos, Avaliação da recuperação da distribuição de covariáveis em relação ao conjunto de dados original (D) - Avaliação do risco de reidentificação de indivíduos a partir dos dados originais, privacidade diferencial e k-anonimato (P)	(SMITH; LAMBERT; RUTHERFORD, 2022)
- Rede Adversarial Gerativa Condicional Diferencialmente Privada (DP-CGANS) - Redes adversárias generativas (GAN) - Redes Bayesianas (BN)	- KL Divergence, Teste Qui-Quadrado de Pearson (D) - Teste de Kolmogorov-Smirnov, Coeficiente V de Cramer, Correlação de Pearson (P)	(SUN; J; M, 2023)

- Rede Adversarial Gerativa de Classificadores Auxiliares (AC-GaN) - Modelo PGaN	- Avaliação da fidelidade da imagem por meio da indistinguibilidade visual de VUs sintéticos e reais, Avaliação da diversidade da amostra gerando um conjunto de dados sintético com localizações variadas de VU (D) - Análise de privacidade à distância de imagens sintéticas dos conjuntos de dados originais usando ataques de pares e de distribuição (P)	(SUN et al., 2023)
- Proposta de um *framework* de dados sintéticos	- Validação cruzada de 10 vezes, Divisão percentual, Matriz de confusão (D)	(UDDIN et al., 2020)
- Imputação múltipla - Substituição iterativa de dados - Análises de regressão de modelos lineares generalizados (GLM) - Análise de componentes independentes (ICA) - Análise de componentes principais (PCA)	- Avaliação da eficiência e viés de dados sintéticos em comparação com dados observados, medição de associações entre preditores e resultados, Mapeamento estatístico paramétrico (SPM), Testes de correlação entre estimativas de informações mútuas observadas e sintéticas, Testes de simulação para avaliar o número de conjuntos de dados sintéticos necessários para uma representação precisa (D)	(VADEN KI et al., 2020)
- Synthea	- Comparação com conjunto de dados real	(WALONOSKI et al., 2018)
- Pré-processamento de k-anonimato - Geração de dados sintéticos diferencialmente privados	- K-anonimato, Método MST, Métodos de pós-processamento para preservar a privacidade (P) - Avaliação da qualidade da inferência, Medição da contagem de regras (D)	(XENIA et al., 2024)
- Criptografia homomórfica (HME) - Privacidade diferencial	- Experiência de classificação, Experiência de consulta de contagem (D)	(JIANG et al., 2018)
- Método de síntese sequencial - Árvores de decisão otimizadas para sequência e com aumento de gradiente - Otimização bayesiana para seleção de hiperparâmetros	- Teste de divulgação de associação (P) - Comparação do modelo de regressão (D)	(AZIZI et al., 2023)

- DataSifterText - MedGan	- Precisão da predição de etiquetas, Acordos médios de predição de rótulos (D) - Avaliações de pontuação BLEU (P)	(ZHOU et al., 2022)
- Rede Adversarial Generativa (GAN) - Codificador automático de rede neural recorrente (RNN-ae) - Autocodificador variacional de rede neural recorrente (RNN-Vae)	- Erro quadrático médio (RMSE), Desvio quadrático médio percentual (PRD) (D)	(ZHU et al., 2019)

Fonte: Elaboração do autor

3.2. Métodos de geração de dados sintéticos e garantia de privacidade

3.2.1. Redes adversariais generativas (GAN) e autocodificadores variacionais (VAE)

Os principais métodos para geração de dados sintéticos encontrados no estudo foram variações de implementação das redes adversariais generativas (GAN) como a condicional (cGAN) (ZH et al., 2024), Wassertein (WGAN)(ARVANITIS et al., 2022) e variações com penalidade de gradiente e/ou normalização espectral (WGAN-GP-SN) (KOSSEN et al., 2021), MedGAN (WALONOSKI et al., 2018), DoppelGanger (DGAN) (ISASA et al., 2024), classificador auxiliar (Ac-GAN)(BEAULIEU-JONES et al., 2019).

Outra metodologia que foi citada entre os estudos é a de geração de dados por autocodificadores variacionais (VAEs). Em dados sintéticos de câncer (RÖCHNER, 2024), foram comparados autocodificadores variacionais (VAEs), redes adversárias generativas (GANs) e modelos probabilísticos de difusão de eliminação de ruído (DDPMs). A análise revelou que os DDPMs produzem dados sintéticos de câncer com mais fidelidade em comparação com aqueles gerados por VAEs e GANs. Apesar disso, os dados sintéticos de câncer produzidos por DDPMs apresentam um risco elevado de privacidade, pois esses dados são mais propensos a divulgar informações relativas a pacientes reais em comparação com os dados sintéticos gerados por VAEs e GANs. Em uma comparação direta com GAN o modelo VAE apresentou performance superior, maior sensibilidade, mas ao mesmo tempo um potencial maior de *overfitting* (LENZ; HESS; BINDER, 2021).

A partir da metodologia de GAN de séries temporais (RTSGAN) os resultados do estudo de (KIM et al., 2024) demonstrou aplicabilidade prática dos dados sintéticos indicando uma semelhança significativa nas distribuições dos conjuntos de dados reais e sintéticos (Hellinger < 0,1), e em relação a utilidade, os dados sintéticos possuem a capacidade de servir como substitutos eficazes para dados reais no contexto do treinamento do modelo (AUC: TSTR e TRTS 0,99 e 0,98, respectivamente).

O estudo de (ISASA et al., 2024) comparou três abordagens distintas de geração de dados sintéticos por WGAN e DGAN: A1 (metadados sintéticos com séries acoplados em tempo real); A2 (criação discreta de metadados sintéticos e séries temporais, seguida por sua posterior fusão); e A3 (geração simultânea de metadados sintéticos e séries temporais e posterior fusão). A avaliação de privacidade revelou que a abordagem A2 foi a que mais preservou a privacidade, enquanto a A3 alcançou as melhores métricas de utilidade. As estratégias A1 e A3 foram identificados como métodos competitivos para gerar dados sintéticos. Em termos de privacidade, o método de redes de produtos de soma mista (MSPNs) demonstrou competência nos critérios de privacidade de classificação como “anônimos” em todos os cenários normais, categóricos e mistos, havendo uma chance mínima de identificação dos pacientes (KROES et al., 2022). Tanto GAN quanto VAE são considerados modelos estados-da-arte, além disso foram observadas outras propostas com menor representatividade entre os artigos selecionados. O estudo de (AZIZI et al., 2023) avaliou um exemplo de geração de dados sintéticos e estudo federado para possibilitar estudos internacionais para saúde cardiovascular. O estudo utilizou uma geração de dados por síntese sequencial por método de árvores e apesar do conjunto de dados ter apresentado alta utilidade e baixo risco de privacidade, a principal limitação ressaltada pela autora foi justamente o uso de apenas uma metodologia para geração de dados.

3.2.2. Privacidade

A utilização de um classificador auxiliar GAN (AC-GAN) diferencialmente privado (DP) para gerar participantes sintéticos para um ensaio clínico, alcançou semelhança estatística e estrutura de correlação dos dados quando comparada as distribuições das variáveis simuladas e reais (BEAULIEU-JONES et al., 2019). Apresentou coerência quando registros individuais foram avaliados por médicos

experientes da área, diferindo em 0.8 pp. entre dados sintéticos e dados reais, portanto o estudo foi teve êxito em representar a pressão arterial sistólica e influência nas decisões de tratamento, embora a introdução da privacidade diferencial tenha adicionado ruído aos dados. A aplicação GANs para anonimizar dados de saúde de pacientes com câncer de pulmão foi relatada pelo estudo de (GONZALEZ-ABRIL et al., 2021) no qual os dados sintéticos gerados correspondiam estreitamente às propriedades estatísticas dos dados reais, alcançando uma precisão discriminatória próxima aos 50% ideais, com uma diferença de 4.66 pp. em relação aos dados reais, reforçando o fenômeno de indistinguibilidade. Apesar disso, limitações foram abordadas quanto a não utilização de todas as variáveis do estudo, que foram selecionadas baseadas em aconselhamento com profissionais clínico, dessa forma pode haver vieses e certas correlações dos dados originais não serão representadas no estudo.

3.2.3. Softwares dedicados

Há também softwares dedicados a geração de dados como Synthea, um simulador populacional que gera dados sintéticos a partir de trabalhos clínicos e informações de progressão da doença, não sendo derivado de dados reais de pacientes (DIOUF et al., 2024). Os resultados de (MEEKER et al., 2022) apontam que o software Synthea pode ser utilizado para investigações de simulação análogas ao estudo de referência sem a necessidade de habilidades avançadas de programação. O estudo de (FORAKER et al., 2020) empregou 3 cenários, trauma pediátrico, predição de sepse e um dashboard de saúde pública, e apesar de conjuntos de dados e abordagens metodológicas variadas, foi verificada confiabilidade na plataforma MDClone sobre a produção de dados com resultados estatísticos equivalentes ou análogos aos dados reais. Em todos os casos de uso, os resultados das análises demonstraram semelhança adequada ($P > 0,05$) entre a derivada sintética e os dados reais, permitindo assim a formulação de conclusões idênticas.

4. DISCUSSÃO

Foi observado que a maior parte dos estudos utilizou variações de redes adversariais generativas e suas variações, assim como diferentes métodos de validações, reforçando a observação de que não há padronização nas métricas. Com o objetivo de discutir as perguntas de pesquisa, serão abordadas as singularidades dos principais métodos, bem como suas características de privacidade e impactos em futuras pesquisas.

4.1. Redes adversariais generativas (GAN)

A proposta da estrutura de redes adversariais generativas foi proposta em 2014 por (GOODFELLOW et al., 2014), se tornando uma arquitetura popular para a geração de dados sintéticos e com modificações subsequentes para atribuição em tarefas específicas (MURTAZA et al., 2023). Inicialmente, a proposta consiste em um “jogo” de minimização de perda máxima para dois jogadores. Consiste em um modelo generativo que captura distribuição de dados e um modelo discriminativo que estima a probabilidade de a amostra ser advinda dos dados de treinamentos ao invés do modelo generativo.

O gerador recebe ruído estocástico como *input*, enquanto produz dados sintéticos como *output*. O discriminador, por sua vez, é equipado com dois *inputs*: os dados de treinamento reais e os dados sintéticos produzidos pelo gerador. A saída fornecida pelo discriminador significa se a entrada é autêntica ou sintética. A tarefa do modelo generativo é de maximizar a probabilidade de o modelo discriminativo cometer um erro. Por outro lado, o modelo discriminativo é treinado para aprimorar sua capacidade de diferenciar entre dados reais e sintéticos (PARK et al., 2018).

O emprego de Redes Adversariais Generativas (GAN) e implementações modificadas foram as metodologias mais prevalentes de geração de dados sintético, sendo observado uma performance superior nas arquiteturas modificadas para atender situações específicas. No estudo de (SUN; J; M, 2023), o emprego de Redes Adversariais Generativas Diferencialmente Privadas e Condicionais (DP-CGANS) gerou resultados superiores a modelos comparáveis, principalmente com respeito a dependência entre variáveis. A modificação impõe uma penalidade adicional para

obrigar o gerador a representar com precisão classes sub-representadas dentro das variáveis desbalanceadas, ao mesmo tempo que simula as correlações e dependências inerentes que existem entre essas variáveis.

Ao envolver a dimensão de tempo há maior complexidade de garantia de tais propriedades estatísticas. Nesse contexto, uma abordagem para gerar dados sintéticos de séries temporais se dá por meio do uso de modelos avançados como o Wasserstein GAN com penalidade de gradiente (WGAN-GP), que incorpora um termo de “perda de alinhamento” para a função geradora, o que garante que as correlações entre as variáveis nos dados sintéticos correspondam às dos dados reais (KUO et al., 2022). Em teoria, o método aumenta a fidelidade dos dados sintéticos e pressupõe a manutenção de padrões e relacionamentos presentes do conjunto de dados original nas séries temporais geradas.

Outro estudo aplicado a doenças cardíacas (ZHU et al., 2019), propôs uma arquitetura inovadora de aprendizado profundo capaz de gerar eletrocardiogramas (ECGs) a partir de conjuntos de dados clínicos, preservando as características intrínsecas dos dados originais. A arquitetura proposta consiste em redes adversárias generativas (GANs), em que uma rede de memória bidirecional de longo prazo (BilsTM) funciona como geradora e uma rede neural convolucional (CNN) serve como discriminadora. O modelo demonstrou produção de ECGs sintéticos que se alinham estreitamente com as distribuições estatísticas presentes nos dados originais de ECG, exibindo características morfológicas que eram análogas aos ECGs autênticos.

Apesar desses resultados, no estudo de (GONCALVES et al., 2020) foram comparados diversos métodos para geração de registros eletrônicos de saúde (EHR), demonstrando que a aplicação de GANs não gerou resultados satisfatórios para o conjunto de dados BREAST (GONCALVES et al., 2020), enquanto outros modelos como *Mixture of Product of Multinomials* (MPoM) and *categorical latent Gaussian process* (CLGP) foram capazes de gerar dados sintéticos com características satisfatórias. Dessa forma, os autores sugerem validações e explorações futuras com a geração de dados por GAN visto a flexibilidade que o modelo proporciona.

4.2. Os Autocodificadores Variacionais (VAEs)

Os autocodificadores variacionais (VAE) empregam redes neurais artificiais (ANNs) para transformar o vetor de entrada em uma representação de menor dimensão dentro do espaço latente, permitindo posteriormente sua reconstrução. A

otimização da rede neural é alcançada por meio da minimização da perda de reconstrução que existe entre a saída e a entrada original. No contexto de um VAE, o codificador mapeia os dados de entrada para uma distribuição de probabilidade caracterizada por seus parâmetros estatísticos, em vez de simplesmente representar um conjunto de vetores de menor dimensão. O processo de aprendizagem do VAE envolve a amostragem dessa distribuição probabilística. Um VAE baseado em trigêmeos melhora a interpretabilidade da representação latente ao integrar um componente adicional de perda de trigêmeos.

No estudo de (ZHU et al., 2019) uma versão uma variante do VAE que emprega um RNN de camada única no codificador e no decodificador, é mencionado como adequado para tarefas discretas, como aprendizado de sequência a sequência e geração de frases e é utilizado como comparador à proposta do estudo.

A variante de VAE chamada *Variational Autoencoder Modular Bayesian Network* (VAMBN) foi desenvolvida para gerar dados longitudinais de estudos clínicos, no estudo de (MOAZEMI et al., 2024) foram empregados dois modelos, VAMBN e MultiNodes para modelar trajetórias longitudinais. O estudo resultou em uma ferramenta web para visualizar e trabalhar com dados de pacientes com Alzheimer. Por outro lado, um outro estudo demonstrou resultados inconclusivos quanto a segurança da privacidade utilizando o mesmo modelo VAMBN (SCHNEIDER et al., 2024).

4.3. Softwares dedicados

O artigo de (MEEKER et al., 2022) explorou o uso do software Synthea abordando a necessidade de criar simulações com uma coorte com as características inerentes à população de Leucemia Mielóide, além disso ressalta que foi necessário funcionalidades específicas que permitissem observar transições de estado complexas e condicionais, que são essenciais para reproduzir com precisão a alocação de tratamento diferencial na coorte de pacientes com Leucemia Mielóide.

Outro sistema de criação de dados sintéticos é o MDClone, o sistema emprega uma ferramenta de consulta para definir populações de estudo e eventos de interesse, que são usados para gerar conjuntos de dados sintéticos estatisticamente semelhantes aos dados reais. O gerador de dados sintéticos do MDClone garante privacidade ao lidar com valores extremos e censurar combinações raras de variáveis para evitar a identificação do paciente (GREENBERG et al., 2023).

4.4. Privacidade

A maior justificativa para utilização de dados sintéticos na saúde é o potencial de acesso sem comprometer a privacidade de pacientes. O vazamento de dados pode ser promovido por ataques em pares, em que uma imagem sintética semelhante a uma amostra de treinamento indica seu uso durante o treinamento, e ataques de distribuição, em que imagens sintéticas se agrupam densamente em torno de imagens reais (SUN et al., 2023). Existem alternativas que utilizam diferentes estruturas para evitar o vazamento de dados. Por exemplo, a privacidade diferencial (DP) é uma estrutura matemática projetada para proteger pontos de dados individuais em um conjunto de dados, garantindo que a inclusão ou exclusão de um único ponto de dados não afete significativamente o resultado de qualquer análise realizada no conjunto de dados. Isso normalmente é obtido adicionando uma quantidade controlada de ruído aos dados ou cálculos, o que obscurece a influência de pontos de dados individuais e, ao mesmo tempo, permite uma análise significativa (BEAULIEU-JONES et al., 2019; LANGE; N; E, 2024). O Departamento do Censo dos EUA adotou a privacidade diferencial para o Censo de 2020, destacando sua aplicação prática na coleta e análise de dados em grande escala (WOOD et al., 2018). Apesar de ser uma proposta com promessa de evitar vazamento de dados, há o custo de adicionar ruído aos dados. O estudo de (SUN et al., 2023) promove uma arquitetura resiliente à ataques por meio do modelo GAN privativo (pGAN), uma abordagem diferente da privacidade diferencial.–A plataforma MDClone emprega uma metodologia que diverge da ocultação de identidades individuais de pacientes. Os modelos são construídos com base em coortes de pacientes que apresentam características semelhantes, gerando novos “pacientes” e a diferenciação entre a coorte sintética e a população real garante a irreversibilidade. O processo de geração de dados sintéticos permite consultas iterativas do conjunto de dados, já que os pacientes sintéticos não são os pacientes originais ofuscados por ruídos (FORAKER et al., 2020)

4.5. Imagens médicas

Uma das maiores potencialidades dos dados sintéticos residem na criação de imagens médicas (LITJENS et al., 2017), um estudo de ablação examinou os efeitos do treinamento de um modelo de desemaranhamento em imagens de peito, pulmão e íris com quantidades variáveis de dados, revelando que modelos treinados com 3.000

e 6.000 imagens poderiam atingir níveis comparáveis de realismo, embora a precisão no reconhecimento de identidade e doença fosse menor com menos dados (MONTENEGRO; CARDOSO, 2024).

Uma das limitações de estudos com imagens médicas é que o processo de sintetizar imagens frequentemente necessita de dados suplementares do paciente, como rótulos de segmentação. Consequentemente, as informações do paciente continuam sendo incorporadas ao modelo, prejudicando a anonimização adequada das imagens geradas. A proposta do estudo de (KOSSEN et al., 2021) foi bem-sucedida em propor a geração combinada de rótulos a imagem sintética, já que para a aplicação das imagens como input de modelos supervisionadas de aprendizado profundo exigem uma anotação manual trabalhosa do conjunto de dados por médicos qualificados. Dessa forma, houve uma anonimização completa com resultados de performance comparáveis aos dados reais.

A técnica de criação de dados rotulados também foi aplicada para imagens de pólipos intestinais (SARAGIH; HIBI; TYRRELL, 2024). Os scores de segmentação desse estudo apresentaram resultados superiores, tanto com conjuntos de dados parcialmente quanto inteiramente sintéticos, quando comparado aos dados reais.

Diferentemente das estratégias citadas anteriormente, o estudo de (VADEN KI et al., 2020) utilizou da técnica de múltipla imputação para o estudo de imagens cerebrais. A técnica consiste em simular diversas iterações do conjunto de dados para gerar informações coletivas de grupos que em última análise refletem os resultados “verdadeiros” observados. Portanto, a estratégia prevê a replicação de resultados estatísticos de grupos, em vez da geração de dados sintéticos “fidedignos” a pacientes individuais. Essa abordagem aponta uma possibilidade de acesso a dados restritos, na qual os investigadores delineiam um modelo de imputação com base nas análises pretendidas, garantindo assim que o conjunto de dados sintéticos resultante apresente utilidade.

4.6. Outras abordagens

Enquanto algumas abordagens tratam de dados tabulares, imagens ou series temporais de eventos médicos, o estudo de (KIM et al., 2024) trouxe a abordagem de síntese simultânea de múltiplas tabelas integradas.

Outro estudo avaliou a validade de dados sintéticos derivados de registros médicos eletrônicos (EMR) comparando-os com dados reais em cinco estudos. Este

estudo demonstrou que dados sintéticos podem fornecer estimativas precisas de resultados de dados reais, apesar de haver perdas considerando conjuntos de dados menores (< 800 pacientes) os resultados apresentaram consistência nos resultados estatísticos, como proporções, razões de chances, razões de risco e curvas de sobrevivência (BENAIM et al., 2020).

O estudo de (RESNICK; COX; MIREL, 2021) foi focado na garantia da privacidade na técnica de *record linkage*, no qual foi comparada a distribuição das estimativas dos parâmetros de sobrevivência entre dados sintéticos e reais, descobrindo que a maioria das estimativas estava dentro de 20% uma da outra, indicando um alto nível de similaridade entre os dois conjuntos de dados.

A discussão sobre as implicações dos dados sintéticos em respeito ao panorama regulatório esteve presente na revisão de (ALLOZA et al., 2023), que elucida os desafios de regulamentações rigorosas no continente Europeu. Em meio a custos substanciais e um tempo considerável, os atrasos na geração de evidências e na obtenção de aprovações regulatórias podem impactar negativamente o atendimento ao paciente. Segundo os autores, os dados sintéticos representam uma alternativa viável para superar esses obstáculos. Iniciativas recentes estão promovendo a incorporação de dados sintéticos em processos de tomada de decisão regulatória e avaliações de tecnologia de saúde; no entanto, os dados sintéticos ainda devem enfrentar desafios práticos antes de serem adotados por pesquisadores e reguladores de forma extensiva na Europa.

5. CONCLUSÃO

O estudo de dados sintéticos é uma primazia tecnológica que tem causado efeitos em diversas áreas do conhecimento e na saúde tem o propósito de viabilizar acesso a dados que não eram possíveis de serem acessados.

A maior parte da literatura elegível para estudo está concentrada a partir de 2020 (n = 43, 81%), demonstrando o caráter recente da aplicação da tecnologia. Os artigos selecionados destacam resultados marcantes em relação aos avanços de arquiteturas para geração de dados sintéticos. O crescente interessa na área médica

e de políticas públicas em saúde pode acelerar o desenvolvimento e popularização de estudos com dados do mundo real (RWD) para fins regulatórios.

Foram abordados diferentes casos de aplicação com dados sintéticos como sensores, eletrocardiograma, textos clínicos, séries temporais, dados administrativos hospitalares, EHR, imagens médicas, simulações farmacocinéticas até a criação de arquiteturas completas para geração de dados sintéticos e estudos de políticas públicas envolvendo a geração de coortes internacionais. Além disso, houve aplicação para um grande espectro de doenças como doenças cardiovasculares, neurodegenerativas e câncer.

Paralelamente, a área carece de padronização de métodos para validação de resultados de performance e de avaliação de segurança, demonstrando uma literatura exploratória, mas ao mesmo tempo com resultados significativos quanto a utilidade dos dados sintéticos. A produção científica na área de geração de dados sintéticos deve garantir a validade e generalização dos dados sintéticos para superar as barreiras à sua adoção por pesquisadores e reguladores. Para além do aperfeiçoamento da técnica, há a necessidade da complacência de estruturas regulatórias mundiais na adaptação ao uso de tais dados, empreendendo esforços contínuos para estabelecer padrões e diretrizes para inovações digitais de saúde baseadas em dados, estabelecimento de critérios de privacidade. No geral, a integração de dados sintéticos na pesquisa e regulamentação da área de saúde está direcionada para transformar o cenário, desde que esses desafios sejam enfrentados por meio de esforços colaborativos entre órgãos reguladores, pesquisadores e partes interessadas do setor médico e farmacêutico.

6. REFERÊNCIAS

ACETO, G.; PERSICO, V.; PESCAPÉ, A. The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges. **Journal of Network and Computer Applications**, v. 107, p. 125–154, 1 abr. 2018.

ALLEN, M.; SALMON, A. **Synthesising artificial patient-level data for Open Science - an evaluation of five methods**. medRxiv, , 13 out. 2020. Disponível em: <https://www.medrxiv.org/content/10.1101/2020.10.09.20210138v1>. Acesso em: 14 set. 2024

ALLOZA, C. et al. A Case for Synthetic Data in Regulatory Decision-Making in Europe. **Clinical pharmacology and therapeutics**, v. 114, n. 4, p. 795–801, out. 2023.

ARORA, A.; ARORA, A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. **Future Healthcare Journal**, v. 9, n. 2, p. 190–193, jul. 2022.

ARVANITIS, T. et al. A method for machine learning generation of realistic synthetic datasets for validating healthcare applications. **Health informatics journal**, v. 28, n. 2, p. 14604582221077000, abr. 2022.

AZIZI, Z. et al. A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. **Scientific reports**, v. 13, n. 1, p. 11540, jul. 2023.

BEAULIEU-JONES, B. et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. **Circulation. Cardiovascular quality and outcomes**, v. 12, n. 7, p. e005122, jul. 2019.

BENAIM, A. R. et al. Analyzing medical research results based on synthetic data and their relation to real data results: Systematic comparison from five observational studies. **JMIR Medical Informatics**, v. 8, n. 2, 2020.

BOSE, S.; MARIJAN, D. **[2311.05404] A Survey on Privacy of Health Data Lifecycle: A Taxonomy, Review, and Future Directions**. Disponível em: <https://arxiv.org/abs/2311.05404>. Acesso em: 12 set. 2024.

CANEDO, E. D. et al. Guidelines adopted by agile teams in privacy requirements elicitation after the Brazilian general data protection law (LGPD) implementation | Requirements Engineering. 2022.

CHEN, J. et al. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. **BMC medical informatics and decision making**, v. 19, n. 1, p. 44, 14 mar. 2019.

D'AMICO, S. et al. Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. **JCO clinical cancer informatics**, v. 7, p. e2300021, jun. 2023.

DIOUF, I. et al. An Approach for Generating Realistic Australian Synthetic Healthcare Data. **Studies in health technology and informatics**, v. 310, p. 820–824, jan. 2024.

DOWNES, E.; HORIGAN, A.; TEIXEIRA, P. The transformation of health care for patients: Information and communication technology, digiceuticals, and digitally enabled care. **Journal of the American Association of Nurse Practitioners**, v. 31, n. 3, p. 156–161, mar. 2019.

EL KABABJI, S. et al. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. **JCO clinical cancer informatics**, v. 7, p. e2300116, set. 2023.

ENANORIA, W. T. A. et al. The Effect of Contact Investigations and Public Health Interventions in the Control and Prevention of Measles Transmission: A Simulation Study. **PLOS ONE**, v. 11, n. 12, p. e0167160, 12 dez. 2016.

ESTEBAN, C.; HYLAND, S. L.; RÄTSCH, G. **Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs**. Disponível em: <https://arxiv.org/abs/1706.02633v2>. Acesso em: 18 set. 2024.

FATEHI, F. et al. General Data Protection Regulation (GDPR) in Healthcare: Hot Topics and Research Fronts. Em: **Digital Personalized Health and Medicine**. [s.l.] IOS Press, 2020. p. 1118–1122.

FERREIRA, L. et al. A panorama of the implementation of the General Law for the Protection of Personal Data (LGPD) in Brazil: an exploratory survey. **2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)**, 2022.

FIGUEIREDO, V. DE B. N.; VARELLA, M. D. DIMENSÕES DA PRIVACIDADE DAS INFORMAÇÕES EM SAÚDE NO BRASIL. **Revista Brasileira de Direitos Fundamentais & Justiça**, v. 17, n. 47, 2022.

FOOD AND DRUG ADMINISTRATION, F. A. D. A. **Real-World Evidence**. Disponível em: <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>. Acesso em: 10 mar. 2024.

FORAKER, R. E. et al. Spot the difference: Comparing results of analyses from real patient data and synthetic derivatives. **JAMIA Open**, v. 3, n. 4, p. 557–566, 2020.

GALLOS, P. et al. INSAFEDARE Project: Innovative Applications of Assessment and Assurance of Data and Synthetic Data for Regulatory Decision Support. **Studies in health technology and informatics**, v. 316, p. 1193–1197, 22 ago. 2024.

GIUFFRÈ, M.; SHUNG, D. L. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. **npj Digital Medicine**, v. 6, n. 1, p. 1–8, 9 out. 2023.

GONCALVES, A. et al. Generation and evaluation of synthetic patient data. **BMC medical research methodology**, v. 20, n. 1, p. 108, maio 2020.

GONZALEZ-ABRIL, L. et al. Generative adversarial networks for anonymized healthcare of lung cancer patients. **Electronics (Switzerland)**, v. 10, n. 18, 2021.

GOODFELLOW, I. J. et al. **Generative Adversarial Networks**. arXiv, , 10 jun. 2014. Disponível em: <http://arxiv.org/abs/1406.2661>. Acesso em: 24 set. 2024

GREENBERG, J. K. et al. Leveraging Artificial Intelligence and Synthetic Data Derivatives for Spine Surgery Research. **Global Spine Journal**, v. 13, n. 8, p. 2409–2421, 2023.

GUAN, J. et al. A Method for Generating Synthetic Electronic Medical Record Text. **IEEE/ACM transactions on computational biology and bioinformatics**, v. 18, n. 1, p. 173–182, fev. 2021.

HAVAEI, M. et al. Conditional generation of medical images via disentangled adversarial inference. **Medical image analysis**, v. 72, p. 102106, ago. 2021.

HAWRYLISZYN, L. O.; COELHO, N. G. S. C.; BARJA, P. R. LEI GERAL DE PROTEÇÃO DE DADOS (LGPD): O DESAFIO DE SUA IMPLANTAÇÃO PARA A SAÚDE. **Revista Univap**, v. 27, n. 54, 26 out. 2021.

HENNESSY, D. Creating a synthetic database for use in microsimulation models to investigate alternative health care financing strategies in Canada. **International Journal of Microsimulation**, v. 8, p. 41–74, 1 nov. 2015.

HITTMEIR, M.; EKELHART, A.; MAYER, R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. **Proceedings of the 14th International Conference on Availability, Reliability and Security**, p. 1–6, 26 ago. 2019.

HUANG, Z. H. et al. Conditional generative adversarial network driven radiomic prediction of mutation status based on magnetic resonance imaging of breast cancer. **Journal of translational medicine**, v. 22, n. 1, p. 226, 2 mar. 2024.

ISASA, I. et al. Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis. **BMC medical informatics and decision making**, v. 24, n. 1, p. 27, jan. 2024.

JANSSEN, A. et al. A Generative and Causal Pharmacokinetic Model for Factor VIII in Hemophilia A: A Machine Learning Framework for Continuous Model Refinement. **Clinical pharmacology and therapeutics**, v. 115, n. 4, p. 881–889, abr. 2024.

JIANG, Y. et al. Privacy-preserving biomedical data dissemination via a hybrid approach. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2018, p. 1176–1185, 2018.

JORDON, J. et al. **Synthetic Data -- what, why and how?** arXiv, , 6 maio 2022. Disponível em: <http://arxiv.org/abs/2205.03257>. Acesso em: 9 mar. 2024

KHAN, S. A.; MURTAZA, H.; AHMED, M. Utility of GAN generated synthetic data for cardiovascular diseases mortality prediction: an experimental study. **Health and Technology**, v. 14, n. 3, p. 557–580, 2024.

KIM, J. et al. Synthesis and quality assessment of combined time-series and static medical data using a real-world time-series generative adversarial network. **Scientific Reports**, v. 14, n. 1, 2024.

KOSSEN, T. et al. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. **Computers in biology and medicine**, v. 131, p. 104254, abr. 2021.

KROES, S. et al. Generating synthetic mixed discrete-continuous health records with mixed sum-product networks. **Journal of the American Medical Informatics Association : JAMIA**, v. 30, n. 1, p. 16–25, dez. 2022.

KUIPER, J.; VAN DEN HEUVEL, E. R.; SWERTZ, M. A. The hybrid synthetic microdata platform: a method for statistical disclosure control. **Biopreservation and biobanking**, v. 13, n. 3, p. 178–182, jun. 2015.

KUO, N. I.-H. et al. The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. **Scientific Data**, v. 9, n. 1, p. 693, 11 nov. 2022.

LANGE, L.; N, W.; E, R. Generating Synthetic Health Sensor Data for Privacy-Preserving Wearable Stress Detection. **Sensors (Basel, Switzerland)**, v. 24, n. 10, maio 2024.

LEE, S. H. Natural language generation for electronic health records. **npj Digital Medicine**, v. 1, n. 1, 2018.

LENZ, S.; HESS, M.; BINDER, H. Deep generative models in DataSHIELD. **BMC medical research methodology**, v. 21, n. 1, p. 64, abr. 2021.

LIBBI, C. A. et al. Generating synthetic training data for supervised de-identification of electronic health records. **Future Internet**, v. 13, n. 5, 2021.

LITJENS, G. et al. A survey on deep learning in medical image analysis. **Medical Image Analysis**, v. 42, p. 60–88, 1 dez. 2017.

LITTLE, C.; ELLIOT, M.; ALLMENDINGER, R. Federated learning for generating synthetic data: a scoping review. **International journal of population data science**, v. 8, n. 1, p. 2158, 2023.

MAMLIN, B. W.; TIERNEY, W. M. The Promise of Information and Communication Technology in Healthcare: Extracting Value From the Chaos. **The American Journal of the Medical Sciences**, v. 351, n. 1, p. 59–68, jan. 2016.

MAMUN, A.-A.; ASELTINE, R.; RAJASEKARAN, S. Efficient Record Linkage Algorithms Using Complete Linkage Clustering. **PloS one**, v. 11, n. 4, p. e0154446, 2016.

MARWALA, T.; FOURNIER-TOMBS, E.; STINCKWICH, S. **The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development**. Disponível em: <https://arxiv.org/abs/2309.00652v1>. Acesso em: 12 set. 2024.

MASOOD, I. et al. Towards Smart Healthcare: Patient Data Privacy and Security in Sensor-Cloud Infrastructure. **Wireless Communications and Mobile Computing**, v. 2018, n. 1, p. 2143897, 2018.

MEEKER, D. et al. Case report: Evaluation of an open-source synthetic data platform for simulation studies. **JAMIA Open**, v. 5, n. 3, 2022.

MOAZEMI, S. et al. NFDI4Health Workflow and Service for Synthetic Data Generation, Assessment and Risk Management. **Studies in health technology and informatics**, v. 317, p. 21–29, ago. 2024.

MOHAMAD, Y. et al. HOW TO OVERCOME LACK OF HEALTH RECORD DATA AND PRIVACY OBSTACLES IN INITIAL PHASES OF MEDICAL DATA ANALYSIS PROJECTS. **Computing and Informatics**, v. 41, n. 1, p. 233–252, 2022.

MONTENEGRO, H.; CARDOSO, J. Anonymizing medical case-based explanations through disentanglement. **Medical image analysis**, v. 95, p. 103209, jul. 2024.

MURTAZA, H. et al. Synthetic data generation: State of the art in health care domain. **Computer Science Review**, v. 48, p. 100546, 1 maio 2023.

NAIK, K. et al. Current Status and Future Directions: The Application of Artificial Intelligence/Machine Learning for Precision Medicine. **Clinical pharmacology and therapeutics**, v. 115, n. 4, p. 673–686, abr. 2024.

NGUFOR, C. et al. Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. **Journal of Biomedical Informatics**, v. 89, p. 56–67, 1 jan. 2019.

NIK, A. H. Z. et al. **Generation of Synthetic Tabular Healthcare Data Using Generative Adversarial Networks**. MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I. **Anais**...Berlin, Heidelberg: Springer-Verlag, 29 mar. 2023. Disponível em: <https://doi.org/10.1007/978-3-031-27077-2_34>. Acesso em: 14 set. 2024

NIKOLENTZOS, G. et al. Synthetic electronic health records generated with variational graph autoencoders. **npj Digital Medicine**, v. 6, n. 1, 2023.

PAIVA, T. DE S.; LANZILLO, A. S. DA S. Proteção de dados pessoais no Brasil: os limites da regulamentação e da regulação da LGPD no constitucionalismo digital brasileiro. **Revista Controle - Doutrina e Artigos**, v. 22, n. 1, p. 239–262, 2024.

PAPROKI, A.; SALVADO, O.; FOOKES, C. Synthetic Data for Deep Learning in Computer Vision & Medical Imaging: A Means to Reduce Data Bias. **ACM Comput. Surv.**, v. 56, n. 11, p. 271:1-271:37, 28 jun. 2024.

PARK, N. et al. Data synthesis based on generative adversarial networks. **Proceedings of the VLDB Endowment**, v. 11, n. 10, p. 1071–1083, jun. 2018.

QUICK, H.; WALLER, L. A. Using spatiotemporal models to generate synthetic data for public use. **Spatial and spatio-temporal epidemiology**, v. 27, p. 37–45, nov. 2018.

RASHIDI, H. et al. Prediction of tuberculosis using an automated machine learning platform for models trained on synthetic data. **Journal of Pathology Informatics**, v. 13, n. 1, p. 10, 2022.

RESNICK, D. M.; COX, C. S.; MIREL, L. B. Using synthetic data to replace linkage derived elements: a case study. **Health Services and Outcomes Research Methodology**, v. 21, n. 3, p. 389–406, 2021.

RÖCHNER, P. On the Fidelity-Privacy Tradeoff of Synthetic Cancer Registry Data. **Studies in health technology and informatics**, v. 316, p. 621–625, ago. 2024.

RUTHERFORD, M. et al. A DICOM dataset for evaluation of medical image de-identification. **Scientific data**, v. 8, n. 1, p. 183, 16 jul. 2021.

SAHI, M. A. et al. Privacy Preservation in e-Healthcare Environments: State of the Art and Future Directions. 2017.

SARAGIH, D.; HIBI, A.; TYRRELL, P. Using diffusion models to generate synthetic labeled data for medical image segmentation. **International journal of computer assisted radiology and surgery**, v. 19, n. 8, p. 1615–1625, ago. 2024.

SCHNEIDER, J. et al. Privacy Risk Assessment for Synthetic Longitudinal Health Data. **Studies in health technology and informatics**, v. 317, p. 270–279, ago. 2024.

SMITH, A.; LAMBERT, P. C.; RUTHERFORD, M. J. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. **BMC medical research methodology**, v. 22, n. 1, p. 176, 23 jun. 2022.

SUN, C.; J, VAN S.; M, D. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. **Journal of biomedical informatics**, v. 143, p. 104404, jul. 2023.

SUN, H. et al. A deep learning approach to private data sharing of medical images using conditional generative adversarial networks (GANs). **PloS one**, v. 18, n. 7, p. e0280316, 2023.

UDDIN, M. A. et al. Rapid health data repository allocation using predictive machine learning. **Health informatics journal**, v. 26, n. 4, p. 3009–3036, dez. 2020.

ULLMAN, J.; VADHAN, S. **PCPs and the hardness of generating private synthetic data**. Proceedings of the 8th conference on Theory of cryptography. **Anais**...: TCC'11.Berlin, Heidelberg: Springer-Verlag, 28 mar. 2011. . Acesso em: 18 set. 2024

VADEN KI, J. et al. Fully synthetic neuroimaging data for replication and exploration. **NeuroImage**, v. 223, p. 117284, dez. 2020.

VALLEVIK, V. B. et al. Can I trust my fake data - A comprehensive quality assessment framework for synthetic tabular data in healthcare. **International Journal of Medical Informatics**, v. 185, p. 105413, maio 2024.

VAN BREUGEL, B. et al. **DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks**. Disponível em: <https://arxiv.org/abs/2110.12884v2>. Acesso em: 18 set. 2024.

WALONOSKI, J. et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. **Journal of the American Medical Informatics Association**, v. 25, n. 3, p. 230–238, 2018.

WOOD, A. et al. Differential Privacy: A Primer for a Non-Technical Audience. **SSRN Electronic Journal**, 2018.

XENIA, H. et al. Studying Privacy Aspects of Learned Knowledge Bases in the Context of Synthetic and Medical Data. **Studies in health technology and informatics**, v. 317, p. 261–269, ago. 2024.

ZH, H. et al. Conditional generative adversarial network driven radiomic prediction of mutation status based on magnetic resonance imaging of breast cancer. **Journal of translational medicine**, v. 22, n. 1, p. 226, mar. 2024.

ZHOU, N. et al. DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes. **Journal of Medical Systems**, v. 46, n. 12, 2022.

ZHU, F. et al. Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network. **Scientific Reports**, v. 9, n. 1, 2019.

Documento assinado digitalmente
VITOR GALVAO LOPES
Data: 11/10/2024 09:47:33-0300
Verifique em https://validar.iti.gov.br

Data e assinatura do aluno(a)

Documento assinado digitalmente
GABRIEL LIMA BARROS DE ARAUJO
Data: 11/10/2024 17:26:51-0300
Verifique em https://validar.iti.gov.br

Data e assinatura do orientador(a)