Redes Bayesianas para Geração de Dados Totalmente Sintéticos para Alta do Recém-Nascido: Uma Proposta Metodológica Orientada a Conhecimento

Jean L. S. Santos[1], **Gilton J. F. da Silva**[1], **Josielson C. da Silva**[2]

Programa de Pós-Graduação em Ciência da Computação
Universidade Federal de Sergipe (UFS)

[2]Programa de Pós-Graduação em Enfermagem e Saúde
Universidade Federal da Bahia (UFBA)

jean.louis.ss@gmail.com, gilton@dcomp.ufs.br,

josielson.silva@ufba.br

Abstract. *Training AI models in neonatology faces data scarcity and privacy constraints. Decision support systems for newborn discharge suffer from the cold start problem. This work proposes a methodology for generating fully synthetic data without relying on real data. The Human-in-the-Loop approach translates clinical guidelines into a Bayesian Network that generates virtual patients. The methodology ensures physiological coherence by implementing deterministic constraints and generating risk classification labels. Preliminary results demonstrate the causal topology of the model, and propose a validation framework comprising visual Turing tests, edit checks, and Machine Learning utility metrics.*

Resumo. *O treinamento de modelos de IA na neonatologia esbarra na escassez de dados e nas restrições de privacidade. Os sistemas de apoio à decisão para a alta de recém-nascidos sofrem com o problema de cold start. Este trabalho propõe uma metodologia para a geração de dados totalmente sintéticos sem o uso de dados reais. A abordagem Human-in-the-loop traduz diretrizes clínicas em uma rede bayesiana geradora de "pacientes virtuais". A metodologia garante a coerência fisiológica implementando restrições determinísticas e gera rótulos de classificação de risco. Os resultados preliminares demonstram a topologia causal do modelo e propõem uma estrutura de validação com testes visuais de Turing, edit checks e métricas de utilidade de Machine Learning.*

1. Introdução e Trabalhos Relacionados

A transição do hospital para o lar é um dos momentos de maior vulnerabilidade na neonatologia: sem critérios padronizados de alta, a decisão fica sujeita à variabilidade do julgamento individual, elevando o risco de reinternações e complicações evitáveis no período pós-alta [Sociedade Brasileira de Pediatria 2020]. Sistemas de apoio à decisão clínica poderiam reduzir essa variabilidade [Chen et al. 2021], mas seu desenvolvimento em neonatologia esbarra em dois obstáculos simultâneos. O primeiro é a escassez de registros clínicos rotulados — o chamado problema do *cold start* —, decorrente da baixa prevalência de eventos adversos e da fragmentação dos prontuários. O segundo é a restrição

regulatória ao compartilhamento de dados sensíveis imposta pela Lei Geral de Proteção de Dados (LGPD) [Gelatti et al. 2021], que limita o acesso mesmo quando os registros existem [Goncalves et al. 2020].

A aplicação de técnicas de aprendizado de máquina (*Machine Learning* - ML) tem demonstrado potencial para a criação de sistemas de apoio à decisão clínica [Chen et al. 2021, Bowe et al. 2023]. Em cenários materno-infantis, redes bayesianas aplicadas à geração de dados sintéticos emergem como alternativa ao cold start, por permitirem representação probabilística compacta e alta explicabilidade, superando as barreiras jurídicas que limitam o compartilhamento de registros reais [Gelatti et al. 2021].

Abordagens *data-driven*, como as *Generative Adversarial Networks* – (GANs) [Choi et al. 2017], alcançam alta fidelidade estatística, mas requerem dados reais como *seed* para aprender as distribuições, mantendo riscos residuais de reidentificação e inferência de pertinência [El Emam et al. 2020]. Abordagens orientadas a conhecimento (*knowledge-driven*), por sua vez, traduzem regras clínicas em simuladores sem depender de registros prévios [Chen et al. 2021].

Diante desse cenário, este trabalho propõe uma abordagem *knowledge-driven* para a geração de dados totalmente sintéticos (*fully synthetic*) voltados à classificação da condição de alta neonatal. Diferente dos métodos *data-driven* que podem gerar ”alucinações”médicas [Goncalves et al. 2020], propomos a modelagem de uma Rede Bayesiana (*Bayesian Network* – BN) estruturada a partir de diretrizes normativas consolidadas [Silva 2024], garantindo plausibilidade fisiológica por meio de restrições determinísticas — os chamados zeros estruturais — e eliminando os riscos de privacidade inerentes ao uso de registros reais [Goncalves et al. 2020].

2. Metodologia Proposta

A metodologia fundamenta-se na geração de dados totalmente sintéticos [Jordon et al. 2022]. Como nenhum dado real é usado no aprendizado da rede — nem mesmo como ponto de partida —, o risco de identificação de pacientes reais a partir dos dados gerados é eliminado por construção [El Emam et al. 2020, Goncalves et al. 2020]. O processo foi estruturado em três etapas principais, detalhadas a seguir.

2.1. Seleção de Variáveis e Modelagem Topológica via Rede Bayesiana

A primeira etapa consistiu na formalização de diretrizes clínicas em uma representação probabilística computacional. O instrumento de referência — validado por enfermeiros e médicos especialistas em neonatologia e desenvolvido a partir de [Silva 2024] — constitui a estrutura normativa de domínio do modelo e organiza 46 variáveis determinantes para a alta segura do recém-nascido em seis eixos: Condições de parto e nascimento, Prontidão biofisiológica, Testagens neonatais, Prontidão e comunicação com os pais, Seguimento de rede de atenção e Condições socioambientais da família. Essa curadoria especializada define o Espaço de Características (*Feature Space*) e atua como *Ground Truth* teórico, garantindo que os dados sintéticos contemplem os aspectos biopsicossociais exigidos nas normativas de alta neonatal [Goncalves et al. 2020].

A seleção das features seguiu abordagem estritamente *Knowledge-Driven*: as relações de dependência entre variáveis foram definidas com base na hierarquia de risco das diretrizes clínicas e no consenso de especialistas — não inferidas por algoritmos

de aprendizado estrutural —, incorporando *prior knowledge* [Acosta et al. 2022] como princípio estruturante. Essa escolha garante a plausibilidade biológica do grafo gerador [Ortega-Leon et al. 2025], reforça a coerência causal entre os seis eixos e mitiga a geração de amostras clinicamente implausíveis.

A modelagem probabilística simultânea de variáveis clínicas categóricas de alta dimensionalidade impõe desafios severos [Goncalves et al. 2020]: o aumento de parâmetros gera esparsidade — a maldição da dimensionalidade [Acosta et al. 2022, Ortega-Leon et al. 2025] —, elevando o risco de *overfitting*. Seguindo [Goncalves et al. 2020], que recomenda isolamento em small-sets para validação preliminar da estabilidade do modelo, as 46 variáveis foram reduzidas a um *Core Set* de 11 atributos (Tabela 1), priorizados pela hierarquia de risco vital [Silva 2024], selecionando as variáveis de maior peso clínico (Pesos 6, 5 e 4) — essenciais para a configuração dos "zeros estruturais" biofisiológicos. Integrou-se também uma variável de baixo risco (Peso 1) — o Esquema vacinal — para validar a capacidade do modelo de aprender dependências entre eixos distintos e relações de causalidade cruzada entre classes de risco heterogêneas.

Tabela 1. Variáveis do *Core Set* priorizadas por peso de impacto clínico

Variável clínica	Eixo de avaliação normativa	Peso
Amamentação	Prontidão biofisiológica	6
Oxigenação	Prontidão biofisiológica	6
Respiração	Prontidão biofisiológica	6
Frequência cardíaca	Prontidão biofisiológica	6
Temperatura corporal	Prontidão biofisiológica	6
Icterícia	Prontidão biofisiológica	6
Peso ao nascer	Prontidão biofisiológica	6
Triagens neonatais	Testagens neonatais	5
Idade gestacional (IG)	Condições de parto e nascimento	4
Apgar no 5º minuto	Condições de parto e nascimento	4
Esquema vacinal	Seguimento de rede	1

Fonte: Elaborado pelo autor.

Para representar as relações probabilísticas entre variáveis clínicas — integrando fatores biofisiológicos, sociais e assistenciais em um modelo graficamente interpretável e clinicamente aplicável [Flores et al. 2023] — modelamos as variáveis por meio de redes bayesianas. O Grafo Direcionado Acíclico (*Directed Acyclic Graph* - DAG) foi desenhado manualmente com base em relações causais validadas por especialistas: a Idade Gestacional (IG) atua como nó pai, influenciando diretamente a distribuição de probabilidade do Peso ao Nascer. Por ser construído exclusivamente por consenso especializado, dispensou-se o uso de algoritmos de aprendizado estrutural e, consequentemente, testes de independência condicional e métricas como p-value, mantendo a coerência da abordagem *knowledge-driven* [Flores et al. 2023]. As Tabelas de Probabilidade Condicional (*Conditional Probability Tables* - CPTs) foram parametrizadas para refletir a frequência esperada de complicações na população.

A implementação utilizou a biblioteca `pomegranate` [Schreiber 2018, Goncalves et al. 2020]. O rótulo final (Alta Segura, Alta com Recomendação ou Alta Contraindicada) não é amostrado probabilisticamente, mas derivado deterministicamente

pela soma ponderada dos scores das variáveis geradas, obedecendo aos limiares clínicos de [Silva 2024], reajustados proporcionalmente ao Core Set (máx. 168 pts).

2.2. Restrições Lógicas e Zeros Estruturais

Para mitigar a geração de combinações clinicamente impossíveis, implementamos *Edit Checks* — restrições determinísticas formuladas como regras *if-then-else* — que definem os zeros estruturais do modelo [Goncalves et al. 2020]. Registros sintéticos que violam consistência cronológica ou biológica — como associar Apgar severamente baixo a "Alta Segura" — são automaticamente rejeitados durante a amostragem.

3. Resultados Preliminares e Estratégia de Validação

A partir da engenharia de conhecimento, 11 dos 46 itens de [Silva 2024] foram estruturados em um DAG funcional (Figura 1), com CPTs parametrizadas por conhecimento de domínio. Dos 1.000 registros sintéticos gerados, o score ponderado derivou deterministicamente: 155 casos de Alta Segura (15,5%), 776 de Alta com Recomendação (77,6%) e 69 de Alta Contraindicada (6,9%).

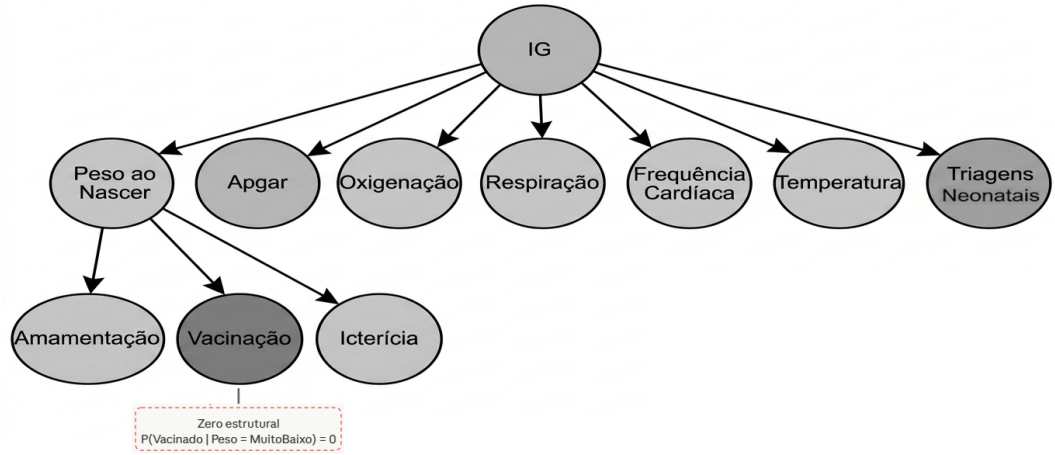

Figura 1. Recorte da topologia causal da Rede Bayesiana.

Como a ausência de dados reais impede a validação por clonagem estatística, a estratégia adotada estrutura-se em dois blocos complementares: a validação do gerador e a validação do instrumento normativo.

3.1. Validação do gerador

O protocolo S1 (Fidelidade Estatística) verificou, por teste qui-quadrado e intervalos de confiança binomiais a 95%, consistência entre distribuições empíricas e CPTs teóricas — 7/7 testes aprovados, incluindo a confirmação do zero estrutural $P(\text{Vacinado} \mid \text{Peso} = \text{MuitoBaixo}) = 0$ com tolerância de 1×10^{-9}. O protocolo S2 (Inferência Condicional) executou sete casos de teste clínicos — 7/7 aprovados, com erros absolutos da ordem de 10^{-16}, confirmando a correção do mecanismo de propagação de crenças. O protocolo S3 (Utilidade Preditiva) treinou um *Random Forest* externo em validação cruzada estratificada de cinco *folds*, reportando Kappa ponderado quadrático $\kappa = 0{,}9178 \pm 0{,}0472$ — classificação *Quase Perfeito* [Landis and Koch 1977] — com ganho de $+0{,}9178$ sobre o *baseline* majoritário. O protocolo S7 (Fidelidade Topológica

Multivariada) avaliou a estrutura probabilística global via Distância de Hellinger e Diferença de Correlação Pareada (PCD): todas as 11 variáveis apresentaram $H < 0{,}10$ (máx.: Temperatura, $H = 0{,}0212$) e PCD MAD sobre os 55 pares de $0{,}0319 < 0{,}05$ [Goncalves et al. 2020, El Emam et al. 2022], confirmando preservação das distribuições marginais e da estrutura de dependência conjunta definida pelas CPTs.

3.2. Validação do instrumento normativo

O protocolo S4 (Curva ROC Sintética) obteve AUC = 1,0000 — resultado esperado na fase PoC, pois o rótulo de classe deriva deterministicamente do *score* contínuo, tornando a separação perfeita por construção matemática. A informatividade clínica de S4 está condicionada à disponibilização de desfechos reais (reinternação, óbito pós-alta) como variável resposta, configurando limitação declarada desta fase. O protocolo S5 (Sensibilidade do Limiar) identificou instabilidade estatística na distribuição de classes (variância de Alta Contraindicada $= 0{,}003022 > 0{,}001$), com região de alta densidade em torno de 144 pts — distante do limiar normativo de 112 pts, sugerindo que este permanece em zona de relativa estabilidade. O protocolo S6 (Casos-Limite Compensatórios) não identificou perfis paradoxais na zona de borda (*score* $\in [110, 114]$): a mediana do *score* biofisiológico nos casos aprovados foi de 95,2% do máximo da categoria ($P_{10} = 81{,}0\%$), confirmando que as dependências causais da BN inibem combinações clinicamente improváveis.

4. Considerações Finais e Próximos Passos

A metodologia propõe uma solução *privacy-by-design* para o *cold start* na alta neonatal: uma rede bayesiana orientada a conhecimento que gera dados totalmente sintéticos sem *seed* real, anulando riscos de reidentificação. O framework de validação — sete protocolos cobrindo fidelidade estatística, inferência condicional, utilidade preditiva, robustez do limiar normativo e fidelidade topológica multivariada (Hellinger + PCD) — demonstra a viabilidade metodológica da abordagem na fase de Prova de Conceito.

Os próximos passos incluem expansão do *Core Set* de 11 para as 46 variáveis de [Silva 2024], incorporando os eixos psicossociais ausentes; TSTR pleno (*Train on Synthetic, Test on Real*), condicionado à disponibilização de dados clínicos institucionais, convertendo S3 e S7 em métricas de validação externa genuína; e Teste Visual de Turing com neonatologistas, complementando a auditoria quantitativa dos Edit Checks com julgamento clínico especializado. Alinhados às práticas de Ciência Aberta [Norori et al. 2021], o código gerador e o dataset sintético serão disponibilizados em repositório público como modelo adaptável para outras especialidades com escassez de dados.

Os autores utilizaram o Google NotebookLM como apoio à síntese e organização da literatura. Todo o conteúdo foi verificado e validado pelos autores, que assumem integral responsabilidade pela acurácia factual, coerência técnica e integridade ética do texto.

Referências

Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. (2022). Multimodal biomedical ai. *Nature Medicine*, 28:1773–1784.

Bowe, A. K., Lightbody, G., Staines, A., Murray, D. M., and Norman, M. (2023). Prediction of 2-year cognitive outcomes in very preterm infants using machine learning methods. *JAMA Network Open*, 6(4):e2349111.

Chen, R., Lu, M., Chen, T., Williamson, D., and Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5:1–5.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare Conference*, pages 286–305. PMLR.

El Emam, K., Mosquera, L., and Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *Journal of Medical Internet Research*, 22(11):e23139.

El Emam, K., Mosquera, L., and Fang, X e El-Hussuna, A. (2022). Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR Medical Informatics*, 10:e35734.

Flores, C., Macagnan, F., Almeida, S., Galante, R., Bykowski, A., Alonso, E., and Oliveira, A. (2023). Sr-bayes: um framework para criação de sistemas de apoio à decisão clínica baseados em redes bayesianas. In *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 186–191, Porto Alegre, RS, Brasil. SBC.

Gelatti, G., Rodrigues, P., and Carvalho, A. C. (2021). Detecção de anomalia através da comparação de modelos representativos. In *Anais Estendidos do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 7–12, Porto Alegre, RS, Brasil. SBC.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20(108).

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. (2022). Synthetic data – what, why and how? *arXiv preprint arXiv:2205.03257*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. (2021). Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2(10):100347.

Ortega-Leon, A., Urda, D., Turias, I. J., Lubián-López, S. P., and Benavente-Fernández, I. (2025). Machine learning techniques for predicting neurodevelopmental impairments in premature infants: a systematic review. *Frontiers in Artificial Intelligence*, 8:1481338.

Schreiber, J. (2018). pomegranate: fast and flexible probabilistic modeling in python. *Journal of Machine Learning Research*, 18(164):1–6.

Silva, C. T. d. S. (2024). *Aplicativo para alta segura do recém-nascido*. Tese de doutorado, Escola de Enfermagem, Universidade Federal da Bahia, Salvador, BA.

Sociedade Brasileira de Pediatria (2020). Recomendações para alta hospitalar do recém-nascido termo potencialmente saudável. Technical Report 7, Departamento Científico de Neonatologia – Sociedade Brasileira de Pediatria. Documento Científico.