Tópico 1: Qualificação da Base de Dados e Treinamento da IA

Resumo Expandido

1. Contextualização

A incorporação de Inteligência Artificial (IA) nos sistemas de saúde brasileiros é uma realidade crescente. Modelos preditivos, sistemas de apoio à decisão clínica, análise de imagens médicas e processamento de linguagem natural são apenas algumas das aplicações que já operam ou se aproximam da escala assistencial. Contudo, a eficácia, a segurança e a equidade dessas ferramentas dependem, de forma absolutamente crítica, da qualidade dos dados que as alimentam.

O Comitê de Inovação e Tecnologia da SBIS reconhece que a qualificação da base de dados representa o alicerce sobre o qual qualquer estratégia de IA em saúde deve ser construída. Sem dados estruturados, padronizados, completos e eticamente obtidos, o resultado dos algoritmos de aprendizado de máquina pode reproduzir — e amplificar — viéses sistemáticos, colocando em risco tanto a qualidade do cuidado quanto a confiança do sistema de saúde.

Este resumo apresenta os cinco sub-tópicos que estruturam a discussão do Encontro 1 do CIT, oferecendo contexto conceitual, pontos de atenção e questões norteadoras para o debate.

2. Sub-Tópicos em Discussão

a) Captura do Dado

A qualidade de qualquer sistema de IA começa no ponto de origem: a captura do dado. Em saúde, os dados são gerados em contextos de alta complexidade — consultas clínicas, internações, exames complementares,
registro multiprofissional, prescrições, sistemas de imagem, dispositivos de monitoramento remoto e relatos do próprio paciente. A heterogeneidade dessas fontes, associada à fragmentação dos sistemas de informação em saúde no Brasil, gera dados incompletos, duplicados e não harmonizados.

Do ponto de vista do treinamento de IA, a captura inadequada gera dois problemas centrais: (i) sub-representação de populações ou condições clínicas, levando a modelos que generalizam mal; e (ii) inserção de ruído nos dados de treinamento, reduzindo a acurácia preditiva e a reprodutibilidade. Estratégias de

captura estruturada — por meio de formulários inteligentes, pontos de entrada padronizados e integridade referencial nos sistemas — são indispensáveis.

Pontos para reflexão:

- Quais são os principais gargalos de captura nos sistemas de saúde públicos e privados brasileiros?
- Como garantir completude sem aumentar a carga administrativa sobre os profissionais de saúde?
- O uso de dados estruturados vs. não estruturados (notas clínicas em linguagem natural) no treinamento de IA.

b) Qualidade do Dado

Qualidade de dado em saúde é um constructo multidimensional. As dimensões mais relevantes para o treinamento de IA incluem: acurácia (o dado representa corretamente a realidade clínica?), completude (campos obrigatórios foram preenchidos?), consistência (os valores são coerentes entre si e ao longo do tempo?), atualidade (o dado reflete o estado atual do paciente?) e unicidade (não há duplicidade de registros?).

A baixa qualidade dos dados é um dos fatores mais citados na literatura como limitante para a escala de IA em saúde. Uma revisão sistemática publicada no JAMIA (Lewis et al., 2023) — que analisou 90 estudos sobre avaliação de qualidade de dados em PEPs publicados entre 2013 e 2023 — concluiu que, apesar do crescimento expressivo de publicações na área, ainda não existe abordagem padronizada para essa avaliação, sendo completude, corretude, concordância, plausibilidade e atualidade as cinco dimensões mais consistentemente investigadas. Em complemento, um scoping review publicado no JMIR Medical Informatics (Penev et al., 2024), que examinou 26 estudos primários, identificou completude (citada em 81% dos estudos), conformidade (69%) e plausibilidade (62%) como os indicadores de qualidade mais frequentemente avaliados — com a maioria dos estudos apresentando limitações de replicabilidade e generalização dos resultados. Para o CIT-SBIS, a discussão sobre qualidade de dado também deve contemplar a definição de padrões mínimos aceitáveis para conjuntos de dados destinados ao treinamento de modelos, assim como mecanismos de auditoria e monitoramento contínuo.

Pontos para reflexão:

- Quais critérios de qualidade devem ser exigidos em um processo de certificação de bases de dados para IA?
- Como a SBIS pode contribuir para a definição de indicadores de qualidade de dados em saúde?
- Papel das ferramentas de Data Quality Assessment (DQA) nos fluxos de governança de dados.

c) Segurança da Informação e LGPD

Os dados de saúde constituem categoria especial de dados pessoais sob a Lei Geral de Proteção de Dados (LGPD, Lei nº 13.709/2018), demandando nível mais rigoroso de proteção. No contexto do treinamento de IA, surgem questões jurídicas e técnicas específicas: qual a base legal para o tratamento massivo de dados de pacientes com finalidade de pesquisa e desenvolvimento de algoritmos? Como garantir o direito à explicação e revisão de decisões automatizadas previsto no art. 20 da LGPD?

Do ponto de vista técnico, técnicas de anonimização e pseudonimização são fundamentais, embora pesquisas recentes demonstrem que dados anonimizados podem ser re-identificados por cruzamento com outras bases. O uso de Privacy-Enhancing Technologies (PETs) como aprendizado federado, computação multipartidária segura e privacidade diferencial surge como caminho promissor para viabilizar o treinamento de IA sem centralização de dados sensíveis.

Pontos para reflexão:

- Como a SBIS pode orientar boas práticas de conformidade com a LGPD para empresas de healthtech que desenvolvem soluções de IA?
- Quais os requisitos mínimos de segurança da informação para bases de dados usadas em treinamento de IA?
- Aprendizado federado como alternativa à centralização: viabilidade no contexto hospitalar brasileiro.

d) Qualificação do Dado: Terminologia, Padronização e Uniformização

A semântica dos dados é tão importante quanto sua estrutura. Para que sistemas de IA possam aprender padrões clínicos a partir de múltiplas instituições, é necessário que os conceitos sejam representados de forma inequívoca e

interoperável. A adoção de terminologias clínicas consolidadas — como SNOMED CT, LOINC, CID-11 — é um pré-requisito para a construção de bases de dados de treinamento de alta qualidade.

No Brasil, a adesão a esses padrões ainda é heterogênea. Sistemas de informação hospitalares frequentemente operam com tabelas internas de códigos, nomes comerciais de medicamentos sem mapeamento para terminologias internacionais, e diagnósticos registrados em texto livre. Esse cenário limita drasticamente a capacidade de treinar modelos generalista e auditável seguros. A SBIS, por seu histórico de promoção de padrões de interoperabilidade, tem papel estratégico na definição de requisitos terminológicos para sistemas que aspirem a desenvolver ou integrar IA clínica.

Pontos para reflexão:

- Qual o papel da SBIS na promoção da adoção de terminologias clínicas padronizadas?
- Como o processo de certificação SBIS pode incentivar ou exigir o uso de SNOMED CT, LOINC e CID-11?
- Iniciativas nacionais para a criação de datasets curados e padronizados para treinamento de IA em saúde.

e) Consentimento do Dado pelo Paciente

O consentimento informado para uso de dados em saúde passa por uma transformação conceitual importante na era da IA. O modelo clássico — estático, por finalidade específica e formalizado em papel no momento do atendimento — mostra-se insuficiente para cobrir o uso dinâmico e multiproposital de dados no contexto do aprendizado de máquina. Dados coletados hoje podem ser necessários para treinar modelos que ainda não existem; pedir novo consentimento para cada uso futuro é inviável operacionalmente. Três modelos alternativos emergem como caminhos possíveis:

O consentimento amplo (broad consent) autoriza, de uma única vez, o uso dos dados do paciente para uma categoria ampla de finalidades — tipicamente "pesquisa em saúde" ou "desenvolvimento de tecnologias assistenciais". É o modelo mais adotado em biobancos e grandes repositórios clínicos, como o UK Biobank. Sua vantagem é a viabilidade operacional; sua limitação é que o paciente assina uma autorização cujo alcance futuro não consegue dimensionar plenamente.

O consentimento dinâmico representa uma evolução: o paciente fornece uma autorização inicial, mas mantém acesso contínuo a uma plataforma digital onde pode, ao longo do tempo, ajustar preferências, autorizar projetos específicos, retirar consentimentos anteriores e ser notificado sobre os usos efetivos de seus dados. Desenvolvido originalmente nas iniciativas CHRIS (Itália) e RUDY (Reino Unido), o modelo pressupõe letramento digital e infraestrutura de governança — requisitos que representam desafio relevante no contexto brasileiro.

O consentimento gerenciado pelo paciente (patient-managed consent) vai além: o paciente detém ativamente a soberania sobre seu conjunto de dados (data sovereignty), decidindo individualmente quem acessa, para qual finalidade e por quanto tempo, com poder de revogação granular. Plataformas como o Personal Health Train (Europa) exploram essa abordagem combinada com aprendizado federado, de modo que os dados podem ser utilizados sem sequer sair do ambiente controlado pelo paciente.

A LGPD reconhece o consentimento como uma das bases legais para o tratamento de dados, mas não a única. Para dados de saúde, o interesse público, a pesquisa científica e a tutela da saúde também figuram como bases legítimas. O desafio prático — e a questão central para o CIT-SBIS — é definir qual modelo de consentimento equilibra adequadamente proteção do paciente e viabilidade operacional no contexto do SUS e da saúde suplementar brasileira, considerando as assimetrias de letramento digital existentes.

Pontos para reflexão:

- Qual modelo de consentimento é mais adequado para o uso de dados de saúde no treinamento de IA em escala?
- Como garantir que paciente de baixo letramento digital possa exercer efetivamente seus direitos previstos na LGPD?
- Plataformas de consentimento dinâmico: experiências internacionais e possibilidades de adoção no Brasil.

3. Questões Norteadoras para o Encontro

Com base nos sub-tópicos apresentados, o CIT propõe as seguintes questões centrais para estruturar a discussão do encontro:

- Quais são os requisitos mínimos que uma base de dados deve atender para ser considerada adequada ao treinamento de sistemas de IA em saúde no Brasil?
- Como a SBIS pode contribuir — por meio de normas, certificações e diretrizes — para a qualificação das bases de dados utilizadas por sistemas de IA clínica?
- Quais mecanismos de governança de dados são necessários para assegurar conformidade com a LGPD sem restringir a inovação?
- Como promover a interoperabilidade semântica nas instituições de saúde brasileiras como pré-condição para IA clínica de qualidade?
- Que modelo de consentimento garante ao mesmo tempo proteção do paciente e viabilidade operacional para pesquisa e desenvolvimento em IA?

4. Referências de Apoio

Os documentos e publicações a seguir constituem leitura recomendada para aprofundamento dos temas:

- BRASIL. Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais – LGPD).
- ANVISA. Regulamentação de Software como Dispositivo Médico (SaDM) – RDC 657/2022.
- WHO. Ethics and Governance of Artificial Intelligence for Health. Genebra: OMS, 2021.
- TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. Nature Med icine, 2019.
- LEWIS, A. E. et al. Electronic health record data quality assessment and tools: a systematic review. Journal of the American Medical Informatics Association (JAMIA), v. 30, n. 10, p. 1730–1740, 2023. DOI: 10.1093/jamia/ocad120. PMID: 37390812.
- PENEV, Y. P. et al. Electronic Health Record Data Quality and Performance Assessments: Scoping Review. JMIR Medical Informatics, v. 12, e58130,

2024. PMID: 39504136. Disponível em: https://medinform.jmir.org/2024/1/e58130

- SBIS. Manual de Certificação para Sistemas de Registro Eletrônico em Saúde. Versão vigente.
- COHEN, I. G. et al. The legal and ethical concerns that arise from using complex predictive analytic s in health care. Health Affairs, 2014.
- HL7 International. FHIR R4 – Fast Healthcare Interoperability Resources. Disponível em: https://hl7.org/fhir/

Documento elaborado pelo CIT-SBIS para circulação prévia aos participantes.

A distribuição é