DIRETRIZES CIENTÍFICAS PARA APLICAÇÃO DE INTELIGÊNCIA ARTIFICIAL EM PESQUISA EM SAÚDE

Laboratório de Neurofisiologia Eduardo Oswaldo Cruz (LNEOC)
Instituto de Ciências Biológicas - Universidade Federal do Pará

"**Qualidade sobre quantidade. Rigor sobre rapidez. Ciência sobre hype**"

APRESENTAÇÃO

Este documento estabelece as diretrizes científicas fundamentais para o desenvolvimento de pesquisas que aplicam inteligência artificial na área de saúde no âmbito do LNEOC. O objetivo é garantir que toda produção científica do laboratório seja fundamentada em rigor metodológico, método científico apropriado e contribua efetivamente para o avanço do conhecimento na área.

A proliferação de estudos que aplicam IA "pela IA" - sem questões de pesquisa bem formuladas, sem hipóteses testáveis e sem contribuição científica clara - representa um dos principais desafios contemporâneos na intersecção entre inteligência artificial e ciências da saúde. Este documento visa equipar os membros do laboratório com ferramentas conceituais e práticas para evitar essas armadilhas e produzir ciência de qualidade.

1. FUNDAMENTOS DO MÉTODO CIENTÍFICO EM PESQUISA COM IA

1.1 A Importância do Método Científico Rigoroso

O método científico permanece como fundamento essencial de qualquer pesquisa de qualidade, independentemente das ferramentas tecnológicas utilizadas. A aplicação de inteligência artificial não substitui nem dispensa o rigor metodológico tradicional.

Elementos essenciais do método científico que DEVEM estar presentes:

- Observação sistemática de fenômenos
- Formulação de perguntas de pesquisa específicas
- Desenvolvimento de hipóteses testáveis
- Design experimental apropriado

- Coleta e análise de dados
- Interpretação de resultados
- Comunicação transparente e reprodutível

1.2 Diferenciação entre Aplicação Técnica e Pesquisa Científica

É fundamental distinguir entre a aplicação técnica de IA e a pesquisa científica que utiliza IA. Nem toda aplicação de IA constitui pesquisa científica.

Aplicação Técnica	Pesquisa Científica
Foco em resolver um problema prático	Foco em responder uma pergunta científica
Não requer hipótese formal	Requer hipótese testável explícita
Sucesso medido por desempenho técnico	Sucesso medido por avanço no conhecimento
Pode usar métodos estabelecidos	Contribui com novos conhecimentos ou validações
Documentação técnica suficiente	Requer publicação científica peer-reviewed

1.3 O Papel da Hipótese Científica em Projetos de IA

Uma hipótese científica é uma afirmação testável sobre a relação entre variáveis. Em pesquisa com IA em saúde, a hipótese deve ir além de "o modelo X pode prever Y".

Exemplos de hipóteses inadequadas vs. adequadas:

❌ INADEQUADO: "Redes neurais profundas podem diagnosticar Parkinson"

✓ ADEQUADO: "Características espectrais de sinais EEG em bandas theta e alfa, quando processadas por CNN, apresentam acurácia superior a 85% na diferenciação entre pacientes com Parkinson em estágio inicial e controles saudáveis, superando métodos baseados em análise temporal convencional"

❌ INADEQUADO: "Machine learning pode melhorar o tratamento de diabetes"

✓ ADEQUADO: "Modelos de gradient boosting que incorporam dados longitudinais de glicemia, atividade física e ingestão alimentar podem prever episódios hipoglicêmicos com até 30 minutos de antecedência com especificidade >90%, permitindo intervenções preventivas"

2. FORMULAÇÃO DE PERGUNTAS DE PESQUISA E HIPÓTESES

2.1 Estrutura PICOT Adaptada para IA em Saúde

A estrutura PICOT (Population, Intervention, Comparison, Outcome, Time) pode ser adaptada para pesquisas que aplicam IA:

P - População: Qual a população-alvo? Quais as características clínicas relevantes?

I - Intervenção/Índice: Qual o modelo/algoritmo de IA? Quais variáveis de entrada?

C - Comparação: Qual o padrão-ouro ou método existente para comparação?

O - Outcome (Desfecho): Qual o desfecho clínico ou variável preditiva de interesse?

T - Tempo: Qual o horizonte temporal? Há follow-up longitudinal?

2.2 Critérios para Perguntas de Pesquisa de Qualidade

Uma pergunta de pesquisa de qualidade deve ser:

F - Feasible (Viável): Existem dados, recursos e expertise disponíveis?

I - Interesting (Interessante): A resposta contribui para o conhecimento na área?

N - Novel (Nova): A pergunta já foi respondida adequadamente?

E - Ethical (Ética): O estudo pode ser conduzido eticamente?

R - Relevant (Relevante): Os resultados terão impacto clínico ou científico?

2.3 Hipóteses Testáveis vs. Demonstrações Técnicas

Uma hipótese científica deve ser falsificável e específica. Deve estabelecer relações claras entre variáveis e prever resultados mensuráveis.

Componentes de uma boa hipótese em IA para saúde:

- Especificação clara do modelo/algoritmo
- Definição precisa das variáveis preditoras
- Desfecho mensurável e clinicamente relevante
- Magnitude esperada do efeito
- Comparação com baseline ou método existente
- Condições e limitações da aplicabilidade

2.4 Alinhamento entre Pergunta, Hipótese e Metodologia

Deve haver coerência lógica entre a pergunta de pesquisa, a hipótese formulada e a metodologia escolhida. Desalinhamentos são indicativos de problemas conceituais no projeto.

3. DELIMITAÇÃO E ESCOPO DE PROJETOS

3.1 Definição Clara de Objetivos Primários e Secundários

Todo projeto deve ter objetivos claramente hierarquizados:

Objetivo Primário: O objetivo principal que justifica o estudo. Deve estar diretamente relacionado à hipótese principal e ser único (ou muito limitado em número).

Objetivos Secundários: Objetivos adicionais que podem ser explorados, mas não são essenciais para o sucesso do projeto. Devem ser claramente identificados como secundários.

IMPORTANTE: Projetos com múltiplos objetivos primários geralmente indicam falta de foco e devem ser subdivididos em projetos separados.

3.2 Critérios de Inclusão e Exclusão

A definição precisa de critérios de inclusão e exclusão é fundamental para:

- Garantir a validade interna do estudo
- Permitir a replicação por outros pesquisadores
- Definir claramente a população à qual os resultados se aplicam
- Evitar vieses de seleção
- Justificar a generalização dos resultados

3.3 Justificativa Científica vs. Justificativa Tecnológica

A justificativa de um projeto de pesquisa deve ser primariamente CIENTÍFICA, não tecnológica.

❌ **Justificativa Inadequada**	✓ **Justificativa Adequada**
"IA é uma tecnologia promissora"	"Métodos atuais apresentam limitação X, e abordagem Y pode superá-la"
"Deep learning tem sido aplicado com sucesso em várias áreas"	"A complexidade não-linear dos dados sugere que modelos lineares são insuficientes"
"Queremos testar o algoritmo X neste problema"	"Hipótese Z ainda não foi testada adequadamente, e modelo X permite este teste"

3.4 Avaliação de Viabilidade e Valor Científico

Antes de iniciar qualquer projeto, avaliar:

- Disponibilidade e qualidade dos dados necessários
- Tamanho amostral suficiente para responder à pergunta
- Expertise técnica disponível no laboratório
- Recursos computacionais necessários

- Tempo realista para conclusão
- Aprovação ética quando necessária
- Potencial de contribuição científica original
- Viabilidade de publicação em periódicos de qualidade

4. QUALIDADE E INTEGRIDADE DE DADOS

4.1 Requisitos de Qualidade de Dados para Modelos de IA

Modelos de IA são fundamentalmente dependentes da qualidade dos dados. O princípio "garbage in, garbage out" é particularmente crítico.

Aspectos críticos de qualidade de dados:

- Completude: proporção de dados ausentes deve ser documentada e justificada
- Acurácia: procedimentos de validação e verificação dos dados
- Consistência: ausência de contradições internas
- Temporalidade: dados atualizados e representativos do período de interesse
- Relevância: variáveis coletadas são pertinentes à pergunta de pesquisa
- Proveniência: origem e histórico de transformações documentados

4.2 Identificação e Mitigação de Vieses nos Dados

Vieses nos dados de treinamento são uma das principais fontes de problemas em modelos de IA aplicados à saúde.

Tipos principais de viés a considerar:

- Viés de seleção: populações sub-representadas ou super-representadas
- Viés de medição: instrumentos ou procedimentos não uniformes
- Viés temporal: mudanças em protocolos ou tecnologias ao longo do tempo
- Viés de confusão: variáveis não controladas que afetam relações
- Viés de label: erros ou inconsistências nas anotações/diagnósticos
- Viés de disponibilidade: dados mais facilmente acessíveis não são representativos

Para cada tipo de viés identificado, deve haver estratégia documentada de mitigação.

4.3 Representatividade de Populações e Generalização

A composição demográfica e clínica dos dados deve ser explicitamente documentada e comparada com a população-alvo de aplicação.

Características a documentar obrigatoriamente:

- Distribuição etária
- Distribuição por sexo/gênero

- Distribuição étnico-racial quando relevante
- Características socioeconômicas quando aplicável
- Comorbidades e status clínico
- Severidade da condição estudada
- Contexto de coleta (hospitalar, ambulatorial, comunitário)

4.4 Gestão de Dados Ausentes e Outliers

Dados ausentes e outliers devem ser identificados, caracterizados e tratados de forma metodologicamente apropriada.

Dados Ausentes:

- Quantificar e reportar a proporção de dados ausentes por variável
- Caracterizar o mecanismo de ausência (MCAR, MAR, MNAR)
- Justificar método de imputação escolhido
- Realizar análises de sensibilidade
- Documentar impacto na interpretação dos resultados

Outliers:

- Definir critérios objetivos para identificação
- Investigar se representam erros ou fenômenos reais
- Documentar decisões de inclusão/exclusão
- Avaliar impacto no desempenho do modelo

4.5 Documentação Completa da Origem e Processamento dos Dados

Toda a trajetória dos dados, desde a coleta até o uso no modelo, deve ser documentada de forma que permita reprodução.

Elementos a documentar:

- Fonte original dos dados (instituição, estudo, base de dados)
- Período de coleta
- Critérios de seleção aplicados
- Todas as transformações realizadas (código disponível)
- Estratégias de normalização e padronização
- Criação de variáveis derivadas
- Versionamento dos datasets

5. DESIGN EXPERIMENTAL E METODOLOGIA

5.1 Seleção Apropriada do Design de Estudo

O design de estudo deve ser escolhido com base na pergunta de pesquisa, não na conveniência ou disponibilidade de dados.

Design	Indicado para	Limitações
Retrospectivo	Estudos exploratórios iniciais, desenvolvimento de modelos	Vieses de seleção, dados incompletos
Prospectivo	Validação de modelos, estudos confirmatórios	Tempo e custo elevados
Cross-sectional	Diagnóstico, classificação em momento único	Não permite inferências causais
Longitudinal	Prognóstico, predição de evolução	Perda de follow-up, complexidade analítica
Caso-controle	Condições raras, estudos de associação	Viés de seleção, recall bias
Coorte	Predição de desfechos, análise de fatores de risco	Custo e tempo consideráveis

5.2 Cálculo e Justificativa do Tamanho Amostral

O tamanho amostral deve ser planejado a priori, não determinado pela disponibilidade de dados.

Considerações para cálculo amostral em ML:

- Regra empírica: mínimo de 10-20 eventos por variável preditora (EPV)
- Para deep learning: regras mais complexas, dependente da arquitetura
- Considerar partição treino/validação/teste (tipicamente 60-70%/15-20%/15-20%)
- Levar em conta desbalanceamento de classes
- Planejar análises de subgrupos e poder estatístico
- Documentar se tamanho amostral é adequado para a complexidade do modelo

ATENÇÃO: Amostras pequenas com modelos complexos são receita para overfitting. Seja conservador na escolha da complexidade do modelo.

5.3 Estratégias de Validação (Interna e Externa)

Validação rigorosa é o que separa um modelo científico de um exercício técnico.

Validação Interna:

- Hold-out: divisão única treino/teste (mínimo aceitável)
- K-fold cross-validation: mais robusto, especialmente para amostras menores
- Leave-one-out: para amostras muito pequenas (custo computacional alto)
- Stratified sampling: manter proporções de classes
- Validação temporal: treino em período anterior, teste em período posterior

Validação Externa: Teste em dados completamente independentes, preferencialmente de instituições ou populações diferentes. Este é o padrão-ouro para demonstrar generalização.

5.4 Evitando Vazamento de Dados entre Conjuntos

Data leakage é uma das causas mais comuns de resultados artificialmente inflados e não-reproduzíveis.

Formas comuns de vazamento a evitar:

- Normalização usando estatísticas de todo o dataset antes da divisão
- Feature selection usando todo o dataset
- Oversampling antes da divisão treino/teste
- Uso de informações temporais futuras
- Inclusão de variáveis proxy do outcome
- Duplicatas não identificadas entre conjuntos

REGRA DE OURO: Qualquer operação que use informações estatísticas dos dados deve ser realizada EXCLUSIVAMENTE no conjunto de treino, nunca envolvendo dados de teste.

5.5 Validação Cruzada Aninhada e suas Aplicações

Para seleção de hiperparâmetros com avaliação não-enviesada de desempenho, validação cruzada aninhada (nested cross-validation) é necessária.

Estrutura: um loop externo para avaliação de desempenho e um loop interno para seleção de hiperparâmetros, garantindo que nenhuma informação do conjunto de teste "vaze" para o processo de otimização.

6. DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS

6.1 Justificativa para Seleção de Algoritmos

A escolha de algoritmos deve ser JUSTIFICADA com base nas características dos dados e na pergunta de pesquisa, não em modismos ou preferências pessoais.

Considerar:

- Natureza dos dados (tabulares, imagens, séries temporais, texto)
- Tamanho da amostra vs. complexidade do modelo
- Necessidade de interpretabilidade
- Relações esperadas (linear vs. não-linear)
- Custo computacional
- Estado da arte para problema similar
- Trade-off between interpretabilidade e performance

Recomendação: começar com modelos mais simples (baseline) e justificar a necessidade de modelos mais complexos com base em ganho de desempenho estatisticamente significativo.

6.2 Métricas de Desempenho Apropriadas ao Contexto Clínico

A escolha de métricas deve refletir as consequências clínicas de diferentes tipos de erro.

Métricas a reportar obrigatoriamente:

- Acurácia (insuficiente sozinha, especialmente em dados desbalanceados)
- Sensibilidade (recall): proporção de casos positivos corretamente identificados
- Especificidade: proporção de casos negativos corretamente identificados
- Valor Preditivo Positivo (precisão): proporção de predições positivas que são corretas
- Valor Preditivo Negativo: proporção de predições negativas que são corretas
- AUC-ROC: discriminação em diferentes thresholds
- Curvas de calibração: concordância entre probabilidades preditas e observadas
- F1-score: média harmônica de precisão e recall

IMPORTANTE: Reportar apenas acurácia é INADEQUADO. Sempre reportar intervalo de confiança ou desvio padrão das métricas.

6.3 Comparação com Padrão-Ouro ou Métodos Existentes

Um novo modelo deve ser comparado com:

- Padrão-ouro clínico quando existente
- Métodos tradicionalmente utilizados para o mesmo problema
- Baselines simples (e.g., regressão logística, árvores de decisão)
- Estado da arte reportado na literatura
- Performance de especialistas humanos quando relevante

Comparações devem incluir testes estatísticos apropriados (e.g., teste de McNemar para comparação de classificadores, teste de DeLong para comparação de AUC-ROC).

6.4 Avaliação de Subgrupos Populacionais

Modelos podem apresentar desempenho heterogêneo em diferentes subgrupos. Isso deve ser investigado sistematicamente.

Subgrupos a avaliar:

- Faixas etárias diferentes
- Sexo/gênero
- Grupos étnico-raciais

- Diferentes níveis de severidade da condição
- Presença/ausência de comorbidades
- Diferentes contextos de aplicação (hospital vs. ambulatório)

6.5 Análise de Curvas de Aprendizado

Curvas de aprendizado (performance vs. tamanho amostral) fornecem insights sobre:

- Se o modelo está underfitting ou overfitting
- Se mais dados melhorariam significativamente a performance
- Se a complexidade do modelo é apropriada
- Qual o tamanho amostral mínimo necessário

7. INTERPRETABILIDADE E EXPLICABILIDADE

7.1 Necessidade de Modelos Interpretáveis em Saúde

Na área de saúde, onde decisões afetam vidas humanas, a capacidade de entender e explicar predições do modelo não é opcional - é essencial.

Razões para exigir interpretabilidade:

- Confiança clínica: profissionais precisam entender o raciocínio
- Validação científica: relações aprendidas devem fazer sentido clínico
- Detecção de vieses: modelos "black box" podem perpetuar vieses ocultos
- Requisitos regulatórios: aprovação por agências de saúde
- Responsabilidade legal: necessidade de justificar decisões
- Educação: modelos devem gerar insights que avancem o conhecimento

7.2 Técnicas de Explicabilidade (SHAP, LIME, etc.)

Quando modelos complexos são necessários, técnicas de explicabilidade devem ser aplicadas:

Método	Tipo	Aplicação
SHAP (SHapley Additive exPlanations)	Global/Local	Importância de features baseada em teoria dos jogos
LIME (Local Interpretable Model-agnostic Explanations)	Local	Aproximação local por modelo interpretável
Attention mechanisms	Local	Visualização de regiões relevantes (imagens, sequências)
Partial Dependence Plots	Global	Efeito marginal de features
Feature importance	Global	Ranking de relevância de variáveis

Counterfactual explanations	Local	O que mudaria a predição?

7.3 Equilíbrio entre Desempenho e Interpretabilidade

Há frequentemente um trade-off entre performance e interpretabilidade. A decisão deve ser consciente e justificada.

Diretrizes:

- Se modelos simples (regressão, árvores) têm desempenho comparável, preferir simplicidade
- Ganhos marginais de performance (<5%) geralmente não justificam perda total de interpretabilidade
- Para aplicações de alto risco, interpretabilidade pode ser prioritária
- Sempre reportar performance de baseline interpretável para comparação
- Documentar explicitamente o trade-off e justificar a escolha

7.4 Comunicação de Resultados para Clínicos

Resultados devem ser comunicados em linguagem acessível a profissionais de saúde:

- Evitar jargão técnico de machine learning desnecessário
- Usar visualizações intuitivas
- Explicar quais informações o modelo usa
- Esclarecer limitações e casos onde o modelo pode falhar
- Fornecer contexto clínico para as predições
- Apresentar incerteza de forma compreensível

8. REPRODUTIBILIDADE E TRANSPARÊNCIA

8.1 Crise de Reprodutibilidade em IA para Saúde

Estudos recentes demonstram que grande parte das pesquisas em IA para saúde não são reproduzíveis, mesmo quando código e dados estão disponíveis. Isso representa uma crise de credibilidade científica.

Principais causas de não-reprodutibilidade:

- Documentação insuficiente de hiperparâmetros
- Falta de controle de aleatoriedade (random seeds)
- Versões diferentes de bibliotecas
- Pré-processamento não documentado
- Data leakage não detectado
- Cherry-picking de resultados
- Múltiplos testes sem correção

8.2 Documentação Completa de Hiperparâmetros

TODOS os hiperparâmetros, inclusive aqueles deixados em valores default, devem ser explicitamente documentados.

Documentar:

- Arquitetura do modelo (layers, units, etc.)
- Algoritmo de otimização e learning rate
- Batch size e número de epochs
- Técnicas de regularização (dropout, L1/L2)
- Data augmentation aplicada
- Método de inicialização de pesos
- Critérios de early stopping
- Estratégia de seleção de hiperparâmetros

8.3 Compartilhamento de Código e Dados (quando possível)

A ciência aberta é o padrão-ouro. Quando não há impedimentos éticos ou legais, código e dados devem ser compartilhados.

Código:

- Repositórios públicos (GitHub, GitLab)
- Código bem documentado e organizado
- Requirements.txt ou environment.yml com dependências
- README com instruções de uso
- Licença apropriada (e.g., MIT, Apache 2.0)

Dados: Quando possível (respeitando ética e privacidade), disponibilizar em repositórios como Zenodo, Figshare, ou específicos da área. Se dados não podem ser compartilhados, compartilhar dados sintéticos que preservem características estatísticas.

8.4 Pré-registro de Estudos

Pré-registro consiste em documentar hipóteses, métodos e análises planejadas ANTES de acessar os dados de teste.

Benefícios:

- Previne HARKing (Hypothesizing After Results are Known)
- Reduz p-hacking e múltiplos testes
- Aumenta credibilidade dos resultados
- Permite distinguir análises confirmatórias de exploratórias

Plataformas: Open Science Framework (OSF), ClinicalTrials.gov (para estudos clínicos).

8.5 Publicação de Resultados Negativos

Resultados negativos (onde a hipótese não foi confirmada) são tão cientificamente valiosos quanto resultados positivos e devem ser publicados.

Importância:

- Evita duplicação de esforços por outros pesquisadores
- Contribui para compreensão mais completa do fenômeno
- Reduz publication bias
- Demonstra rigor científico do laboratório

9. DIRETRIZES DE REPORTE (REPORTING GUIDELINES)

Diretrizes de reporte (reporting guidelines) são checklists de itens que devem ser incluídos em publicações para garantir transparência e reprodutibilidade. Seu uso é OBRIGATÓRIO no LNEOC.

9.1 TRIPOD+AI para Modelos Preditivos

TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + AI) é a extensão do TRIPOD para modelos de IA.

Aplicação: estudos que desenvolvem, validam ou atualizam modelos preditivos.

Principais domínios: título/abstract, introdução, métodos (participantes, outcome, preditores, tamanho amostral, dados ausentes, análise estatística, performance), resultados, discussão, informações suplementares.

Disponível em: https://www.tripod-statement.org/

9.2 CONSORT-AI para Ensaios Clínicos

CONSORT-AI é a extensão do CONSORT (Consolidated Standards of Reporting Trials) para ensaios clínicos randomizados que avaliam intervenções baseadas em IA.

Aplicação: ensaios clínicos randomizados testando intervenções com IA.

Itens adicionais específicos para IA incluem: descrição do algoritmo, dados de treinamento, análise de erros, interação humano-IA.

9.3 SPIRIT-AI para Protocolos de Ensaios

SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials - AI extension) para protocolos de ensaios clínicos com IA.

Aplicação: protocolos de ensaios clínicos ANTES de sua condução.

Uso: quando submeter protocolos de pesquisa para aprovação ética ou registro de ensaios.

9.4 STARD-AI para Estudos de Acurácia Diagnóstica

STARD-AI (Standards for Reporting Diagnostic accuracy studies - AI extension) para estudos que avaliam acurácia de testes diagnósticos baseados em IA.

Aplicação: estudos de validação diagnóstica.

9.5 FUTURE-AI: Princípios de IA Confiável

FUTURE-AI é um guideline internacional para desenvolvimento e reporte de IA médica confiável, com foco em aspectos práticos de implementação.

Pilares FUTURE:

- Fairness (Justiça): equidade entre grupos
- Universality (Universalidade): generalização
- Traceability (Rastreabilidade): auditabilidade
- Usability (Usabilidade): integração clínica
- Robustness (Robustez): performance consistente
- Explainability (Explicabilidade): interpretabilidade

Recomendação: consultar FUTURE-AI para aspectos de implementação prática além do reporte de pesquisa.

10. CONSIDERAÇÕES ÉTICAS

10.1 Consentimento Informado e Uso de Dados

O uso de dados de saúde para desenvolvimento de modelos de IA requer consideração ética cuidadosa, mesmo quando usando dados retrospectivos.

Questões a considerar:

- Pacientes consentiram que seus dados fossem usados para esta finalidade?
- É necessário novo consentimento ou há waiver apropriado?
- Dados foram adequadamente anonimizados/de-identificados?
- Há risco de re-identificação, especialmente em amostras pequenas?
- Aprovação do Comitê de Ética em Pesquisa foi obtida?

10.2 Privacidade e Segurança de Dados de Pacientes

Medidas obrigatórias:

- Armazenamento seguro com controle de acesso
- Criptografia em repouso e em trânsito
- Minimização de dados: coletar apenas o necessário
- Logs de acesso aos dados
- Conformidade com LGPD (Lei Geral de Proteção de Dados)
- Plano de resposta a incidentes de segurança
- Destruição segura ao final do estudo

10.3 Equidade e Justiça no Desenvolvimento de IA

Modelos de IA podem perpetuar ou amplificar desigualdades em saúde. É responsabilidade dos pesquisadores identificar e mitigar essas questões.

Avaliar:

- Representação equitativa de populações nos dados de treinamento
- Performance diferencial entre grupos demográficos
- Acesso potencial à tecnologia desenvolvida
- Impacto em disparidades de saúde existentes
- Quem se beneficia e quem pode ser prejudicado?

10.4 Accountability e Responsabilidade

Quando modelos de IA são aplicados em decisões clínicas, questões de responsabilidade emergem.

Considerar:

- Quem é responsável quando o modelo erra?
- Como o modelo será monitorado após deployment?
- Há mecanismos de feedback e atualização?
- Profissionais têm autonomia para discordar do modelo?
- Há documentação clara de limitações?

10.5 Avaliação de Riscos e Benefícios

Análise de risco-benefício deve ser explícita:

Potenciais Benefícios:

- Melhoria em acurácia diagnóstica
- Detecção precoce de condições
- Personalização de tratamentos
- Redução de custos
- Aumento de acesso a cuidados

Potenciais Riscos:

- Erros diagnósticos
- Perpetuação de vieses
- Erosão da relação médico-paciente
- Violação de privacidade
- Dependência excessiva de automação
- Amplificação de desigualdades

11. CHECKLIST DE AVALIAÇÃO DE PROJETOS

11.1 Checklist para Revisão de Propostas

Use este checklist para avaliar propostas de projetos ANTES de iniciar o trabalho:

□ A PERGUNTA DE PESQUISA está claramente formulada?

□ A pergunta segue estrutura PICOT ou similar?

□ Há uma HIPÓTESE testável e específica?

□ A hipótese vai além de "modelo X pode prever Y"?

□ OBJETIVOS primários e secundários estão claramente hierarquizados?

□ Há JUSTIFICATIVA científica (não apenas tecnológica)?

□ O gap de conhecimento está claramente identificado?

□ DADOS necessários estão disponíveis ou há plano de coleta?

□ TAMANHO AMOSTRAL é adequado para a complexidade do modelo?

□ Há cálculo ou justificativa do tamanho amostral?

□ DESIGN de estudo é apropriado para a pergunta?

□ Estratégia de VALIDAÇÃO está clara (interna e externa)?

□ Há plano para evitar DATA LEAKAGE?

□ VIESES potenciais foram identificados?

□ Há estratégias de mitigação de vieses?

□ MÉTRICAS de avaliação são apropriadas ao contexto clínico?

□ Há COMPARAÇÃO com baseline ou padrão-ouro?

□ INTERPRETABILIDADE foi considerada?

□ Há plano para explicar predições do modelo?

□ ASPECTOS ÉTICOS foram abordados?

□ Aprovação do CEP será necessária? Já foi obtida?

□ Há RECURSOS (computacionais, tempo, expertise) disponíveis?

□ VIABILIDADE de conclusão no prazo é realista?

□ Há potencial de PUBLICAÇÃO em periódico de qualidade?

□ REPORTING GUIDELINE apropriado foi identificado?

11.2 Critérios de Qualidade Metodológica

Projetos serão avaliados nos seguintes critérios:

Critério	Excelente	Inadequado
Pergunta de Pesquisa	Específica, PICOT estruturada, lacuna clara	Vaga, ampla demais, sem estrutura
Hipótese	Testável, específica, com magnitude esperada	Ausente ou não testável
Justificativa	Científica, gap claro, referências atuais	Apenas tecnológica, sem contexto
Metodologia	Rigorosa, validação adequada, sem leakage	Validação inadequada, potencial leakage
Tamanho Amostral	Calculado/justificado, adequado	Não justificado, inadequado
Interpretabilidade	Estratégia clara de explicabilidade	Modelo black-box sem justificativa
Reprodutibilidade	Código disponível, documentação completa	Documentação insuficiente

11.3 Pontos Críticos de Validação

Revisões em marcos do projeto devem focar nestes pontos críticos:

- Após formulação: pergunta e hipótese são científicas?
- Após coleta: dados têm qualidade suficiente?
- Após pré-processamento: há risco de leakage?
- Após desenvolvimento: validação é rigorosa?
- Após resultados iniciais: interpretação está correta?
- Antes da submissão: reporting guideline foi seguido?

12. ERROS COMUNS A EVITAR

12.1 "IA pela IA": Aplicação sem Pergunta Científica

ERRO: "Vamos aplicar deep learning neste conjunto de dados e ver o que descobrimos"

CORRETO: "Hipótese: características não-lineares de imagens de ressonância podem predizer resposta a tratamento melhor que critérios radiológicos convencionais. Vamos testar isso com CNNs"

Por que é um erro: Sem pergunta científica, não há como julgar sucesso, escolher metodologia apropriada, ou contribuir para conhecimento. É exploração de dados, não pesquisa.

12.2 Tamanho Amostral Inadequado

ERRO: "Temos 100 pacientes, vamos treinar uma rede neural com 10.000 parâmetros"

CORRETO: "Com 100 pacientes, podemos estimar um modelo com 5-10 variáveis. Vamos usar regressão logística regularizada ou random forest limitado"

Por que é um erro: Modelos complexos com amostras pequenas levam a overfitting severo. Resultados não são generalizáveis e tendem a não replicar.

12.3 Validação Inadequada ou Ausente

ERRO: "Nosso modelo tem 99% de acurácia no conjunto de treino!"

CORRETO: "Performance no treino: 92%, validação cruzada: 85%, teste externo: 81%. Intervalos de confiança: [77-85%]"

Por que é um erro: Performance no conjunto de treino não significa nada. O que importa é performance em dados nunca vistos, preferencialmente de fontes independentes.

12.4 Desconsideração de Vieses

ERRO: "Coletamos todos os dados disponíveis no hospital e treinamos o modelo"

CORRETO: "Identificamos que pacientes mais graves têm mais exames registrados (viés de observação). Estratificamos por severidade e avaliamos performance em cada estrato"

Por que é um erro: Dados "disponíveis" raramente são representativos. Vieses não identificados levam a modelos que funcionam mal no mundo real.

12.5 Falta de Interpretabilidade Clínica

ERRO: "O modelo funciona, não precisamos entender como"

CORRETO: "Análises SHAP revelam que as variáveis mais importantes são X, Y e Z, o que está alinhado com fisiopatologia conhecida. Identificamos um outlier onde o modelo usou artefato técnico - corrigimos o pipeline"

Por que é um erro: Sem interpretabilidade, não podemos validar se o modelo aprendeu relações causais ou apenas correlações espúrias. Dificulta detecção de erros e impede geração de insights científicos.

12.6 Generalização Excessiva dos Resultados

ERRO: "Nosso modelo diagnostica câncer de pulmão" (testado apenas em um hospital, pacientes >60 anos, um tipo de scanner)

CORRETO: "Nosso modelo discrimina nódulos benignos vs. malignos em TC de tórax, validado em pacientes >60 anos no Hospital X usando scanner Y. Generalização para outras populações e equipamentos requer validação adicional"

Por que é um erro: Modelos de IA são sensíveis a mudanças em populações, protocolos e equipamentos. Afirmações amplas sem validação apropriada são científica e eticamente irresponsáveis.

CONSIDERAÇÕES FINAIS

A aplicação de inteligência artificial em pesquisa em saúde representa uma fronteira empolgante do conhecimento científico. No entanto, o potencial transformador dessas tecnologias só será realizado se fundamentado em rigor metodológico inabalável.

Este documento estabelece as expectativas mínimas para pesquisa no LNEOC. Não são sugestões - são requisitos. A reputação científica do laboratório, a credibilidade de nossos pesquisadores, e o potencial de impacto real em saúde dependem da qualidade metodológica de nosso trabalho.

Lembre-se: estamos formando cientistas, não técnicos em IA. A diferença fundamental está na capacidade de fazer perguntas importantes, formular hipóteses testáveis, desenhar estudos rigorosos, interpretar resultados criticamente, e comunicar descobertas de forma responsável.

Direção do LNEOC

Laboratório de Neurofisiologia Eduardo Oswaldo Cruz

Instituto de Ciências Biológicas - UFPA