CLEDON — Modelo de projeto

Nome de apresentação: CLEDON — Rede colaborativa de dados e inteligência artificial para pesquisa oncológica.

Versão: 1.3 • 15 de novembro de 2026

O que mudou em relação à 1.2: a regra de que dados de pacientes só são processados por modelos executados dentro das instituições, sem envio a serviços externos de inteligência artificial (seção 7, "IA própria e local"); esclarecimento sobre o que a criptografia protege durante o processamento (seção 6). Também nesta edição, depois de revisão técnica: definição de "local", critérios de aceitação para a extração de texto, papel do teste externo, infraestrutura sem exigência universal de GPU e responsável técnico pelos modelos em cada instituição. O demonstrador continua no protocolo 0.2, somente com dados fictícios.

Estágio: estudos com dados públicos concluídos e depositados (Fases 1 a 2b, cledon.org); demonstrador da rede construído com dados fictícios (cledon.org/demonstrador). Para a rede com dados clínicos, esta é uma proposta conceitual: nenhuma parceria, base de pacientes, validação clínica ou eficácia está sendo afirmada.

Em uma página

Pergunta do piloto. Em pacientes com carcinoma epidermoide de esôfago operados após quimiorradioterapia neoadjuvante, dados clínicos e laboratoriais disponíveis antes do tratamento preveem a resposta patológica completa, com desempenho mantido em outro hospital?

O que o piloto entrega. Um estudo retrospectivo em duas instituições de desenvolvimento e uma de teste externo, sem centralizar prontuários: cada hospital mantém seus registros e executa apenas análises aprovadas. Entregas: relatório de viabilidade e qualidade dos dados; modelo comparado a um modelo simples, com discriminação e calibração no teste externo; descrição de como exposições como tabagismo, álcool e bebidas muito quentes estão registradas.

O que pedimos a um parceiro. Indicação de um responsável clínico e de dados; uma contagem agregada de casos elegíveis (quantos iniciaram neoadjuvância e quantos foram operados); participação no desenho do protocolo. Nenhum prontuário é solicitado nesta etapa.

O que já existe. Estudos exploratórios com dados públicos sobre o excesso persistente de mortalidade por câncer de esôfago em regiões do Sul, de Minas Gerais e da Bahia (Zenodo, DOI 10.5281/zenodo.24195057; resumo no Anexo A) e um demonstrador da rede com dados fictícios (seção 11).

1. Resumo executivo

Desenvolver uma infraestrutura que permita a hospitais e centros de pesquisa investigar, de forma coordenada, relações entre características dos pacientes, exames, tratamentos e resultados clínicos, mantendo os registros individuais sob a guarda das instituições participantes.

O projeto começa com uma pergunta oncológica delimitada e amplia sua cobertura conforme demonstre qualidade dos dados, utilidade científica, segurança e sustentabilidade. A ambição de longo prazo é contribuir para novas estratégias terapêuticas e possíveis curas; a entrega inicial é uma rede de pesquisa confiável e um estudo reproduzível.

O diferencial proposto combina integração acessível a instituições com poucos recursos, avaliação em populações locais e governança que limite usos discriminatórios de resultados individuais e territoriais. Trata-se de uma hipótese de diferenciação a verificar, não de uma alegação de ineditismo.

2. O que aprimorar na ideia original

Elemento	Modelo proposto	Benefício esperado, sujeito a avaliação
Cobertura universal desde o início	Começar com um câncer, uma pergunta e poucas instituições	Reduzir dependências e permitir medir resultados
Um banco central com todos os prontuários	Rede federada: bases locais, catálogo central de metadados e estudos autorizados	Reduzir movimentação de dados individuais
Uma IA especial para tudo	Métodos específicos para padronizar dados, testar hipóteses e prever um desfecho definido	Facilitar interpretação e comparação de desempenho
Dados processados por serviços de IA de terceiros	Modelos executados dentro de cada instituição, sob guarda local, sem envio de dados de pacientes para fora	Reduzir dependência de fornecedores e de jurisdições estrangeiras
Coletar todos os exames disponíveis	Coletar apenas variáveis necessárias para cada protocolo	Reduzir custo, exposição e trabalho de curadoria
Criptografia como proteção principal	Segurança técnica, limites de acesso, avaliação das saídas e governança independente	Cobrir riscos que a criptografia isolada não resolve
Prometer a descoberta de cura	Produzir hipóteses e evidências com uma trajetória de validação	Tornar a proposta cientificamente avaliável

Redes de dados e aprendizado federado já existem. Owkin e colaboradores publicaram pesquisa com IA e dados de hospitais para prever resposta à quimioterapia; o NCI mantém uma infraestrutura de dados oncológicos. Esses precedentes devem constar da candidatura. Estudo na Nature Medicine, NCI CRDC.

3. Primeiro estudo: resposta ao tratamento do câncer de esôfago

Por que esôfago. Os estudos com dados públicos do Cledon encontraram excesso persistente de mortalidade por câncer de esôfago em municípios do Sul, de Minas Gerais e da Bahia. Começar a rede pelo mesmo câncer dá continuidade a esse trabalho e aponta parceiros naturais: hospitais que atendem essas regiões. A rede não pretende explicar com prontuários as causas do excesso regional; isso exigiria outro desenho (por exemplo, caso-controle), com protocolo e aprovação próprios.

Pergunta proposta: em pacientes com carcinoma epidermoide de esôfago operados após quimiorradioterapia neoadjuvante, dados clínicos e laboratoriais disponíveis antes do tratamento conseguem prever a resposta patológica completa, com desempenho que se mantenha em outro hospital?

População-alvo da previsão: pacientes operados após a neoadjuvância. A resposta patológica só existe para quem chega à cirurgia; por isso, o desempenho do modelo vale para esse grupo e não deve ser apresentado como válido para todos os pacientes que iniciam a neoadjuvância. Quem não chega à cirurgia é um grupo selecionado (por progressão, óbito, condição clínica ou recusa), e registrá-lo no fluxograma é necessário, mas não elimina essa seleção. Uma previsão com desempenho demonstrado para todos os que iniciam o tratamento exigiria outro desfecho, definido para todos (por exemplo, chegar à cirurgia com resposta completa, ou sobrevida), a decidir com os oncologistas.

A quimiorradioterapia neoadjuvante seguida de cirurgia é um tratamento estabelecido para o câncer de esôfago localmente avançado; no ensaio CROSS, cerca de metade dos pacientes com carcinoma epidermoide teve resposta patológica completa. van Hagen et al., NEJM 2012. A resposta patológica completa é definida pelo exame da peça cirúrgica; neste projeto, seria um desfecho de pesquisa, não sinônimo de cura nem substituto automático da sobrevivência. O protocolo deve fixar a definição operacional com o oncologista e o patologista antes da análise.

Objetivo secundário (descritivo): descrever como as exposições associadas na literatura a esse câncer (tabagismo, álcool, bebidas muito quentes como o chimarrão, cachaça) estão documentadas nos prontuários de cada hospital. É uma avaliação da qualidade e da viabilidade do registro, não um estudo de risco: sem grupo de comparação, essas contagens não indicam se uma exposição aumenta o risco.

Configuração proposta: duas instituições para desenvolvimento e uma terceira reservada para teste externo, de preferência em regiões diferentes (por exemplo, Sul, Minas Gerais e Bahia). São metas de parceria, não instituições confirmadas. Se só houver uma ou duas, a entrega será de viabilidade ou validação limitada, sem alegação de generalização ampla.

Desenho inicial: estudo retrospectivo, usando registros já existentes e autorizados. Incluir todos os inícios de tratamento elegíveis no período definido; documentar quem não chegou à cirurgia (progressão, óbito, condição clínica ou recusa), teve desfecho desconhecido ou abandonou o acompanhamento. Quem não foi operado não tem resposta patológica avaliável e não pode ser contado como "sem resposta". Não converter ausência de registro em ausência de doença ou de exposição.

Recorte clínico: adultos com carcinoma epidermoide de esôfago não metastático que iniciaram quimiorradioterapia neoadjuvante com intenção cirúrgica; critérios finais, regimes e períodos serão definidos no protocolo. O adenocarcinoma, com biologia e resposta diferentes, fica fora do primeiro estudo. Não misturar instituições, épocas e regimes sem examinar essas diferenças.

Risco principal de viabilidade: muitos pacientes são diagnosticados em estágio avançado e não chegam à cirurgia, o que pode deixar poucos casos com o desfecho. A etapa de viabilidade deve contar, de forma agregada, quantos casos elegíveis cada hospital tem. Se forem insuficientes, o desfecho será revisto com os oncologistas (por exemplo, resposta clínica ou sobrevida) ou o trabalho se limitará à viabilidade.

Tamanho da amostra: calculado pelo bioestatístico separadamente para o desenvolvimento e para o teste externo. No desenvolvimento, conforme o número de parâmetros candidatos, a frequência do desfecho e o risco de sobreajuste (Riley et al., BMJ 2020); no teste externo, conforme a precisão desejada da calibração (observado/esperado e inclinação) e da discriminação (Riley et al., BMJ 2024). Não fixar um número de pacientes apenas para tornar a proposta atraente.

Uso inicial: exclusivamente pesquisa, sem recomendar ou alterar tratamento de pacientes.

4. Dados mínimos e qualidade

Grupo	Exemplos propostos	Regra
Contexto clínico	Idade, condições prévias e estado funcional documentado	Apenas campos necessários e com definição consistente
Caracterização do tumor	Estadiamento clínico (cTNM), histologia, localização e biomarcadores disponíveis	Preservar origem, método e momento da medida
Exposições documentadas	Tabagismo, álcool, bebidas muito quentes (como o chimarrão), cachaça	Registrar como documentado; campo vazio não significa "não exposto"
Exames pré-tratamento	Resultados laboratoriais selecionados pelo protocolo	Padronizar unidades e referências; não presumir equivalência
Tratamento planejado	Regime conhecido na data de previsão	Separar intenção de tratamento efetivamente recebido
Evolução	Tratamento realizado, mudanças, eventos adversos e cirurgia	Dados de acompanhamento, nunca entradas retrospectivas indevidas do preditor inicial
Resultado	Resposta patológica definida e perdas de acompanhamento	Revisar uma amostra de prontuários para verificar o rótulo
Contexto territorial	Região ampla ou indicadores de acesso, se justificados	Área restrita; sem endereço preciso no ambiente analítico

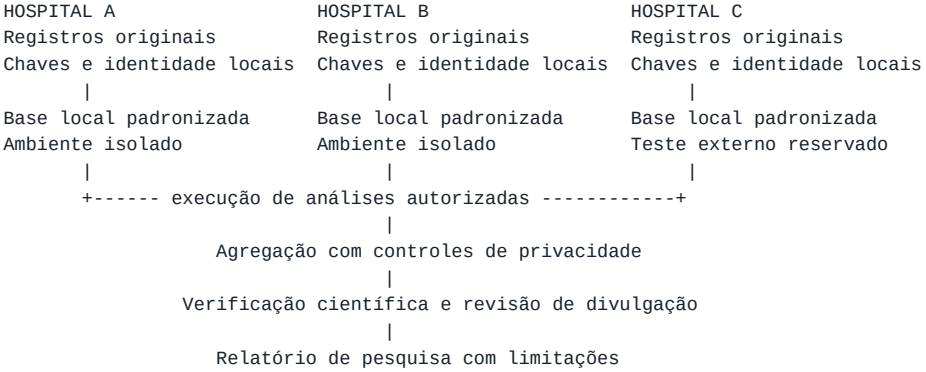
A data da previsão precisa estar explícita. Informações que só se tornam conhecidas depois dessa data não podem entrar no treinamento como se já fossem disponíveis.

Começar com dados estruturados. Texto livre, imagens e genômica entram posteriormente, somente se sua inclusão responder a uma pergunta definida e justificar o custo adicional. Relatórios e imagens podem conter identificadores e exigiriam controles específicos.

Usar um padrão comum de pesquisa, como o **OMOP CDM**, em um subconjunto bem documentado. Manter códigos originais, origem dos campos e registros das transformações. Não afirmar compatibilidade completa sem verificá-la. O OMOP oferece uma estrutura comum para padronizar dados observacionais. Documentação oficial.

O relatório de qualidade deve medir campos ausentes, consistência de datas, duplicidade local, unidades incompatíveis, validade do desfecho e diferenças entre instituições. Possíveis pacientes presentes em mais de uma instituição exigem um procedimento autorizado de identificação de sobreposição; não exportar CPF ou um hash simples de CPF como solução improvisada.

5. Arquitetura funcional



O coordenador central mantém protocolos, permissões, versões de código, catálogo de variáveis disponíveis e resultados autorizados. Não mantém uma cópia universal dos prontuários.

Cada estudo envia um pacote de análise versionado e aprovado para os ambientes locais. O pacote não pode executar consultas arbitrárias ou exportar arquivos. Cada instituição pode recusar ou suspender a execução.

No início, priorizar análises descritivas distribuídas e comparação de modelos locais. Implementar treinamento federado apenas quando necessário para a pergunta e depois de avaliar segurança, custo e capacidade dos parceiros. Dados reais não entram no demonstrador inicial: usar dados fictícios criados para testar o fluxo.

6. Privacidade e proteção contra discriminação

Proteção dos registros: criptografia no armazenamento e na transmissão, autenticação multifator, acesso mínimo necessário, chaves controladas pelas instituições, separação entre identificação e dados analíticos, auditoria, backups e plano de incidentes. Pseudonimização não será anunciada como anonimização irreversível.

Proteção do processamento: ambientes isolados, execução de código aprovado, restrições de rede e exportação, credenciais por estudo e revisão das atualizações de modelo. Aprendizado federado pode vazar informações por atualizações ou modelos; não é garantia isolada de privacidade. NIST — riscos em aprendizado federado.

O que a criptografia cobre: a criptografia protege os dados armazenados e em trânsito. Para serem analisados, em geral os dados precisam ser abertos; por isso, a abertura acontece somente dentro do ambiente isolado da instituição guardiã. Técnicas de computação que protegem as entradas durante o cálculo existem (criptografia homomórfica, que opera sobre dados cifrados; computação segura multipartidária, que opera sobre partes secretas distribuídas), mas são custosas e se aplicam a operações específicas; serão consideradas onde o modelo de ameaças justificar, como na agregação segura de atualizações entre instituições. A proposta não afirma que os dados permanecem protegidos por criptografia durante toda a análise.

Proteção dos resultados: limitar consultas repetidas e combinações que permitam reconstruir registros; avaliar supressão de grupos pequenos e supressão complementar; revisar granularidade territorial; avaliar privacidade diferencial com orçamento e impacto na utilidade documentados. Nenhum limite fixo de tamanho de grupo será tratado como garantia universal.

Agregação segura e privacidade diferencial devem ser escolhidas a partir de um modelo explícito de ameaças. Com poucas instituições, colusão, diferenças entre rodadas e contribuição dominante de um hospital precisam ser consideradas. Não prometer que “nem o administrador consegue saber nada” sem demonstrar as condições técnicas.

Proteção de comunidades: a privacidade de uma pessoa não elimina o risco de discriminação contra uma região. Mesmo estatísticas verdadeiras e agregadas podem ser usadas contra comunidades. Por isso, mapas e resultados territoriais passam por revisão de finalidade e de divulgação; pesquisas de planejamento em saúde podem ter acesso restrito quando justificado.

Regras propostas de uso: vedar contratualmente e operacionalmente precificação individual de seguros, seleção de beneficiários, decisões de contratação de pessoas e marketing baseado em condição clínica. Não disponibilizar API pública de risco por pessoa, CEP ou bairro. Patrocinadores recebem apenas produtos de pesquisa autorizados, sem acesso automático aos registros, consultas ou modelos.

Essas medidas reduzem riscos e criam responsabilização. Não permitem garantir ausência absoluta de vazamentos ou de uso indevido após divulgação.

Antes de qualquer dado real: definir finalidade, base legal aplicável a dados sensíveis, responsabilidades, avaliação ética cabível, retenção, regras de compartilhamento e direitos dos participantes. A participação de universidade ou hospital não torna qualquer uso automaticamente permitido. ANPD — orientação para estudos e pesquisas.

7. IA e demonstração de eficiência

O primeiro comparador será um modelo estatístico simples, definido antes do teste. Uma abordagem mais complexa só permanece se demonstrar ganho relevante, considerando custo e interpretabilidade.

Na primeira etapa, usar métodos adequados a tabelas clínicas.

IA própria e local

Regra: dados de pacientes só são processados por modelos executados dentro do ambiente da instituição guardiã. Nenhum dado de paciente, nem trecho de prontuário, é enviado a serviços externos de inteligência artificial, qualquer que seja o país ou a empresa fornecedora.

O que "local" significa no piloto: infraestrutura nas dependências da instituição guardiã e administrada por ela. Uma nuvem de terceiros, mesmo contratada pelo hospital, fica fora do piloto. É uma restrição inicial deliberada; o comitê de acesso pode revê-la depois, com avaliação específica de segurança, contrato e localização dos dados.

O que pode sair da instituição: prontuários e registros individuais não saem. Resultados agregados saem depois da revisão de divulgação. Atualizações federadas e pesos de modelos podem revelar informações dos registros usados no treino; por isso só saem em estudos autorizados e depois de avaliação própria, separada da avaliação dos resultados agregados.

Camada	Uso	Como fica sob controle
Modelos para dados tabulares	Prever o desfecho do protocolo a partir de variáveis estruturadas (por exemplo, regressão logística, como no demonstrador)	Código do projeto, aberto e versionado, executado em cada instituição; só saem resultados agregados aprovados
Modelos de linguagem para texto livre	Extrair do texto do prontuário variáveis definidas no protocolo (por exemplo, tabagismo, álcool, temperatura do chimarrão), se os campos estruturados não bastarem	Modelos de pesos abertos, baixados uma vez e executados sem conexão de saída para a internet, no servidor da instituição; ajuste local (por exemplo, LoRA) somente se a avaliação justificar
Assistentes externos de programação e redação	Escrever código, documentação e análises com dados públicos	Nunca recebem dados de pacientes; o código produzido é revisado e testado antes de ir às instituições

Por que não treinar um modelo do zero: um modelo de linguagem de grande porte exige investimento, equipe e volume de texto fora do alcance de um projeto de pesquisa. A independência pretendida vem de três escolhas: executar os modelos localmente, manter os pesos sob guarda das instituições e avaliar o desempenho com dados próprios. Modelos de pesos abertos com bom desempenho em português, incluindo iniciativas brasileiras, serão comparados antes da escolha. Uma participação futura em esforços nacionais de modelos em português pode ser considerada, sem ser pré-requisito.

Requisitos antes de usar um modelo de linguagem com dados reais:

- Inventário de modelos: nome, versão, origem, licença e impressão digital (hash) dos pesos; comandos (prompts), configurações e versão do programa de execução.
- Ambiente sem saída para a internet e com registro de auditoria.
- Avaliação contra uma revisão manual de prontuários, numa amostra independente da que foi usada para ajustar os comandos, com métricas por variável e por instituição e relato de erros.
- Limites mínimos de desempenho definidos no protocolo antes da avaliação. Variável que não atingir o limite não entra em análises.
- Cada informação extraída conserva a origem (documento e trecho) e a data do registro.

- O modelo registra o que está escrito, sem deduzir além disso: "consome chimarrão" não autoriza preencher "temperatura acima de 65 °C". Informação ausente continua como desconhecida.
- Relato do estudo conforme a diretriz TRIPOD-LLM. Gallifant et al., Nature Medicine 2025.
- Revisão humana das saídas e regras de supressão iguais às dos demais resultados.
- Pesos ajustados com dados de pacientes podem memorizar informações: são tratados como dados sensíveis e não saem da instituição sem avaliação específica.

Infraestrutura: modelos tabulares rodam em servidores comuns. Modelos de linguagem podem exigir placa gráfica (GPU), conforme o modelo, o volume de texto e o tempo de processamento aceitável; modelos menores ou quantizados também rodam em processador comum (CPU), o que importa para instituições com poucos recursos. A infraestrutura é orçada com cada instituição depois do inventário de dados; nenhum valor é fixado nesta proposta.

Critérios propostos para avançar

Dimensão	O que medir	Decisão
Qualidade	Completude, validade dos desfechos e concordância da revisão clínica	Não treinar quando o desfecho ou os campos essenciais não forem confiáveis
Capacidade preditiva	Discriminação, calibração e incerteza, incluindo comparação com o modelo simples	Escolher e ajustar o modelo só nos hospitais de desenvolvimento. O teste externo avalia o modelo congelado e o comparador, ambos definidos antes; não serve para escolher nem reajustar
Generalização	Desempenho em outro hospital e em período posterior	Não extrapolar para hospitais, regimes ou períodos não avaliados
Equidade	Erros e calibração por subgrupos relevantes e por instituição	Lacunas importantes exigem investigação, limitação de uso ou novo estudo
Privacidade	Testes de acesso indevido, exportação, inferência e revisão das saídas	Falha relevante impede uso ou divulgação até correção
Eficiência operacional	Horas de integração, curadoria por registro, custo por análise e tempo de execução	Comparar com o processo manual e simplificar quando a automação não compensar
Reprodutibilidade	Mesma versão de dados e código reproduz o resultado	Resultados sem trilha verificável não entram em divulgação científica

Os limites numéricos de aprovação serão registrados no protocolo antes do teste, com justificativa clínica e estatística. Não escolher uma “acurácia de 95%” arbitrária como meta.

Reservar pacientes e hospitais para validação; separar períodos; procurar sobreposição entre bases. A escolha e o ajuste do modelo acontecem somente nos hospitais de desenvolvimento, com validação interna. Depois, todo o processo é congelado, incluindo a extração de variáveis por IA, quando houver, e só então é avaliado no hospital externo, uma única vez. O resultado externo é relatado como sair: se o modelo congelado não superar o comparador simples, isso é o resultado, e o modelo não é reajustado para o mesmo teste. Um modelo escolhido ou ajustado com base no teste externo não pode ter esse desempenho apresentado como confirmação independente. Collins et al., BMJ 2024. Relatar intervalos de confiança e perdas de acompanhamento. Pequenos subgrupos sem precisão suficiente devem ser descritos como inconclusivos, não como equivalentes.

Uma boa previsão não demonstra que mudar o medicamento produzirá melhor resultado. Comparações de tratamentos exigem desenhos específicos, controle de confundimento e posterior validação adequada. A literatura do NCI destaca representatividade e validação clínica como desafios centrais. NCI — IA e câncer.

8. Caminho até novas terapias

1. **Infraestrutura:** comprovar integração e qualidade de dados, com governança funcional.
2. **Pesquisa observacional:** identificar padrões de resposta, toxicidade e diferenças entre populações; produzir hipóteses, incluindo resultados negativos.
3. **Replicação:** testar os achados em dados e equipes independentes.
4. **Pesquisa causal e biológica:** avaliar hipóteses em protocolos apropriados; associar laboratórios e especialistas quando necessário.
5. **Validação clínica:** conduzir estudos e ensaios adequados ao uso proposto antes de alterar a assistência.
6. **Expansão:** incorporar outras instituições, modalidades de dados e doenças quando os resultados justificarem.

Não é possível fixar prazo ou probabilidade de encontrar uma cura a partir deste modelo. O avanço concreto é medido pelas etapas de evidência concluídas.

9. Governança e sustentabilidade

Arranjo inicial recomendado: um consórcio de pesquisa com uma instituição científica e hospitais parceiros; o idealizador atua como articulador, enquanto os responsáveis clínicos e científicos conduzem os protocolos de sua área. Nenhuma instituição está confirmada.

Papéis necessários: coordenação oncológica, patologia, bioestatística, engenharia de dados, ciência de dados, segurança e privacidade, responsável local por cada base, responsável técnico pelos modelos em cada instituição e representantes de pacientes. Algumas funções podem ser compartilhadas ou prestadas por parceiros, com responsabilidades explícitas.

Responsável técnico pelos modelos locais: em cada instituição, uma pessoa aprova as versões instaladas, controla atualizações, acompanha erros e pode suspender a execução. A validação das variáveis clínicas extraídas continua sob responsabilidade clínica.

Comitê de acesso: pesquisadores, instituições guardiãs e representantes de pacientes avaliam estudos e divulgação. Patrocinadores declaram conflitos e não decidem sozinhos sobre acesso, análise ou publicação. Registrar decisões e publicar resumos compreensíveis dos projetos.

Financiamento proposto: editais e apoio filantrópico para a prova de conceito; posteriormente, serviços de integração, manutenção e execução de estudos aprovados para instituições. Não basear o modelo de receita na venda de prontuários. Remuneração por serviços não elimina obrigações sobre dados nem conflitos de interesse.

Custos a orçar: horas de curadoria e integração por instituição, equipe científica e técnica, suporte local, infraestrutura, auditoria, governança, avaliação ética e disseminação. Formular orçamento após inventário e propostas dos parceiros; evitar estimativa financeira apresentada como cotação sem evidência.

10. Plano inicial de execução

Os prazos abaixo são hipóteses de planejamento. A fase com dados reais só começa depois dos acordos e aprovações necessários; atrasos institucionais deslocam o cronograma.

Etapa	Janela proposta	Entrega e condição para seguir
Definição	Semanas 1–2	Resumo, comparação com iniciativas existentes, pergunta clínica e responsabilidades

Etapa	Janela proposta	Entrega e condição para seguir
Viabilidade	Semanas 3–6	Conversas com parceiros, inventário de campos e contagens agregadas autorizadas, cálculo preliminar de amostra
Demonstrador	Concluído em 13/11/2026	Três hospitais simulados, registros fictícios, análise distribuída, controle de acesso e revisão de saídas (cledon.org/demonstrador)
Preparação do piloto	Semanas 11–12	Protocolo, orçamento por etapa, plano de segurança e documentação institucional
Piloto com dados reais	Estimativa de 6–12 meses após condições de início	Integração, curadoria, desenvolvimento e teste externo, sujeitos a revisão de escopo

O demonstrador deve mostrar o fluxo completo e os limites de acesso. Dados fictícios demonstram funcionamento do software, não eficácia clínica, distribuição real de pacientes ou segurança definitiva.

Não avançar ao piloto sem parceiro clínico, dados viáveis, cálculo de amostra, responsabilidades, proteção e aprovações aplicáveis. Se não houver desfecho confiável, a entrega será a infraestrutura e o estudo de qualidade dos dados; não produzir um preditor artificialmente convincente.

11. Demonstrador para eventos

Quatro telas bastam para uma primeira demonstração:

1. **Rede participante:** hospitais fictícios e qualidade das bases, sem lista de pacientes.
2. **Protocolo do estudo:** pergunta, finalidade, variáveis aprovadas, responsáveis e versão.
3. **Execução controlada:** análise autorizada executada nos três ambientes simulados e tentativa de exportação bloqueada.
4. **Resultado científico:** comparação com modelo simples, incerteza, limitações e status da revisão de divulgação.

Situação: as quatro telas estão implementadas no demonstrador (cledon.org/demonstrador, com o código aberto), com o piloto de esôfago simulado. Cada hospital simulado confere a aprovação do protocolo pela impressão digital SHA-256, recusa operações e variáveis fora do protocolo, suprime células com menos de 5 pessoas e registra todos os pedidos em auditoria local. O modelo de regressão logística é ajustado de forma federada, com resultado idêntico ao de juntar as bases (verificado em teste), congelado e registrado no hospital reservado para teste externo, que recusa treino e aceita avaliar cada modelo congelado uma única vez. Campos vazios ou inválidos viram ausência. Faltantes e observados são suprimidos em conjunto para impedir a recuperação de contagens pequenas pelo cruzamento dessas duas margens com o total. Outras combinações de saídas, inclusive casos completos de modelos, ainda exigem revisão de divulgação; essa regra não é uma garantia formal de privacidade. O protocolo 0.2 fixa os casos da comparação e os parâmetros de avaliação antes do teste, e a aprovação local impede substituir um modelo já registrado sob o mesmo protocolo.

Marcar todas as telas como “DEMONSTRAÇÃO COM DADOS FICTÍCIOS”. Não mostrar porcentagens inventadas como evidência de sucesso. Um mapa detalhado de risco por bairro não faz parte do demonstrador.

12. Texto-base para candidatura

O CLEDON propõe uma infraestrutura colaborativa de pesquisa oncológica que conecta instituições de saúde para analisar dados clínicos protegidos, sem centralizar os prontuários individuais. O projeto já publicou estudos exploratórios com dados públicos sobre o excesso persistente de mortalidade por câncer de esôfago em regiões do Sul, de Minas Gerais e da Bahia (Zenodo, DOI 10.5281/zenodo.24195057) e construiu um demonstrador da rede com dados fictícios. O primeiro piloto investigará a previsão de resposta à quimiorradioterapia neoadjuvante no carcinoma epidermoide de esôfago, com dados padronizados, modelos executados dentro das próprias instituições (sem envio de dados de pacientes a serviços externos de inteligência artificial), comparação com métodos simples e validação em instituição independente. A proposta dará atenção à participação de hospitais com recursos limitados e à governança contra usos discriminatórios, incluindo revisão de resultados territoriais. Para a rede com dados clínicos, o projeto busca parceiros científicos, instituições participantes e financiamento para viabilizar um estudo retrospectivo. Seu objetivo de longo prazo é gerar evidências e hipóteses que contribuam para tratamentos mais eficazes e para a pesquisa de possíveis curas.

Pedido inicial aos parceiros: indicação de responsável clínico e de dados, avaliação de viabilidade dos registros, participação no desenho do protocolo e orçamento de integração. Não solicitar o envio de prontuários na primeira abordagem.

Próxima decisão: obter revisão independente dos estudos com dados públicos e validar com uma equipe de oncologia que trate câncer de esôfago se a pergunta e o desfecho escolhidos são relevantes e registráveis nos hospitais disponíveis, incluindo quantos pacientes chegam à cirurgia. A partir disso, fechar o protocolo antes de uma candidatura.

Anexo A. Estudos já publicados com dados públicos

Importante: os resultados abaixo são de estudos exploratórios, ainda não revisados por especialistas independentes. Não são conclusões, não indicam causa e não permitem estimar o risco de uma pessoa específica.

Antes de buscar dados de hospitais, o Cledon analisou dados públicos e agregados: 7,9 milhões de declarações de óbito de todas as causas do SIM/DATASUS (2020–2024), das quais 1,2 milhão com câncer como causa básica, e a população do Censo 2022 (IBGE), nos 5.570 municípios. Os óbitos foram atribuídos ao município de residência. Cada município foi comparado com o número de mortes por câncer esperado para sua idade e sexo (padronização indireta), com suavização bayesiana empírica e correção para múltiplas comparações dentro de cada grupo de análise.

Foram identificados 145 sinais estatísticos em 13 grupos de análise. Aplicando critérios geográficos exploratórios (separar capitais, municípios a até 40 km delas e polos de tratamento), 45 permanecem. Esses critérios não demonstram que os vieses foram removidos e podem excluir excessos reais. O câncer de esôfago concentra 16 desses sinais: 5 no Sul, onde estudos já associam esse câncer ao consumo de bebidas muito quentes (como o chimarrão em temperatura muito alta), ao álcool e ao tabaco; 9 em diferentes regiões de Minas Gerais e 2 na Bahia, com mortalidade observada entre 2,4 e 3,6 vezes o esperado pela média nacional após a suavização. Em testes de sensibilidade (outros raios, outros critérios de polo, correção somando os 13 grupos de análise e população anual), 14 dos 16 mantêm o sinal. O excesso aparece em 2020–2022 e em 2023–2024, mas isso é uma checagem de estabilidade, não uma validação independente.

Em 2015–2019, período não utilizado na seleção dos sinais, 29 dos 45 pares município–grupo após os critérios geográficos e 12 dos 16 referentes ao câncer de esôfago apresentaram novamente excesso de mortalidade pelo critério definido (plano de análise registrado localmente, no git, antes do exame dos dados). Esses resultados sustentam a persistência temporal dos padrões observados, mas não distinguem excessos reais de vieses persistentes de registro, de endereço ou de diagnóstico, nem estabelecem causalidade.

Os estudos seguintes, cada um com plano de análise registrado localmente antes dos dados, aprofundaram o câncer de esôfago:

- **Fase 1b:** no Registro Hospitalar de Câncer do INCA, uma fonte diferente das declarações de óbito, os 16 municípios têm participação do câncer de esôfago entre os casos de câncer acima da esperada para o estado (14 atendem ao critério estatístico); uma varredura espacial encontrou 24 áreas regionais com excesso de mortalidade.
- **Fase 1c:** com óbitos de 1996 a 2014, que não participaram da seleção, os 16 municípios juntos tiveram em 2000–2014 mortalidade por câncer de esôfago 1,91 vez a esperada para seus estados (IC 95% 1,74–2,10), resultado mantido com território constante; pulmão e outros cânceres ficaram abaixo do esperado. Parte da alta ao longo do tempo pode refletir a melhora do registro das causas de morte.
- **Fases 2 e 2b (geração de hipóteses):** no país, o excesso acompanha água de fonte, nascente ou poço, extração de erva-mate, população rural, lenha e fumo; no Sul, erva-mate e fumo; em Minas Gerais, ruralidade, menor PIB per capita, menor altitude e maior produção rural de cachaça. São associações ecológicas entre municípios, não causas.

A inteligência artificial foi usada como assistente de programação e redação, sob revisão humana; por ser um serviço externo, recebeu somente dados públicos e agregados. Nenhum modelo de aprendizado de máquina foi treinado ou avaliado nesses estudos. O relatório técnico completo, com fontes, versões, parâmetros e código, está em cledon.org/relatorio e foi depositado no Zenodo, em português e em inglês (versão 1.7: DOI 10.5281/zenodo.24195057; todas as versões: https://doi.org/10.5281/zenodo.23093761). O primeiro pedido aos parceiros científicos é uma revisão independente desses estudos.

Anexo B. Homenagem e identidade

Cledon é uma homenagem a Francisco Cledio Capistrano, pai do idealizador Clériston Capistrano, que faleceu em decorrência do câncer. O slogan **Find Cancer Clues** preserva suas iniciais F.C.C. O projeto está em formação, sem alegação de eficácia clínica.