Pipeline Reprodutível e Responsável de Ciência de Dados para Apoio à Detecção Precoce do Câncer de Pulmão

Douglas L. P. Olinto, Luciana O. Berretta, Elisângela S. Dias

[1]Instituto de Informática, Universidade Federal de Goiás

douglas.olinto@discente.ufg.br

luciana.berretta@ufg.br

elisangeladias@ufg.br

Abstract. *This paper presents a reproducible data-engineering pipeline for public-health-oriented data science in lung oncology. The workflow structures public biomedical sources through provenance registration, integrity validation, quality control, semantic label extraction, deterministic matrix–metadata alignment, leakage-aware partitioning, and versioned artifacts. In the audited execution, GSE43458 and GSE32863 were consolidated with complete binary labels and no unclassified samples, which can contribute to reducing technical barriers to reproducible research by documenting provenance, labeling rules, data-quality checks, and versioned artifacts.*

Resumo. *Este artigo apresenta um pipeline reprodutível de engenharia de dados para Ciência de Dados orientada à saúde pública em oncologia pulmonar. O fluxo estrutura fontes biomédicas públicas por meio de proveniência, validação de integridade, controle de qualidade, extração semântica de rótulos, alinhamento matriz–metadados, particionamento anti-vazamento e artefatos versionados. Na execução auditada, GSE43458 e GSE32863 foram consolidadas com rótulos binários completos e sem amostras não classificadas, o que pode contribuir para reduzir barreiras técnicas à pesquisa reprodutível ao documentar proveniência, regras de rotulagem, verificações de qualidade e artefatos versionados.*

1. Introdução

O câncer de pulmão permanece como um dos principais desafios de saúde pública, tanto pela elevada carga de morbimortalidade quanto pela necessidade de estratégias mais eficazes de diagnóstico precoce, estratificação de risco e acompanhamento populacional [Siegel et al. 2024, World Health Organization 2024]. Paralelamente, o uso de dados biomédicos públicos e de técnicas de inteligência artificial (IA) tem ampliado as possibilidades de pesquisa em oncologia, especialmente em contextos nos quais a coleta primária de dados clínicos é limitada, onerosa ou indisponível [Beam and Kohane 2018].

Apesar desse avanço, a construção de evidências computacionais em saúde ainda enfrenta problemas recorrentes de rastreabilidade, heterogeneidade de metadados, inconsistência de rótulos, substituição silenciosa de arquivos, baixa documentação de transformações e risco de vazamento entre treino e teste. Diretrizes como TRIPOD (*Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis*) e

PROBAST (*Prediction model Risk Of Bias ASsessment Tool*) reforçam que desenho de dados, transparência metodológica e avaliação de risco de viés fazem parte da evidência científica, não apenas da implementação [Collins et al. 2015, Wolff et al. 2019].

Nesse cenário, a dificuldade investigada não está somente na escolha do algoritmo, mas na cadeia de dados que antecede qualquer modelagem. Arquivos públicos heterogêneos precisam ser localizados, verificados, rotulados, alinhados, filtrados e versionados antes que possam sustentar análises reprodutíveis. Essa etapa é particularmente crítica em oncologia pulmonar, pois diferentes domínios biológicos, como tecido tumoral, TEP (*tumor-educated platelets*) e cfRNA (*cell-free RNA*), não devem compartilhar rótulos, partições ou métricas sem protocolo explícito.

Este trabalho propõe um *pipeline* reprodutível de engenharia de dados para integração, curadoria e preparação de bases públicas voltadas à análise de câncer de pulmão. Neste artigo, *pipeline* designa um fluxo computacional encadeado de ingestão, validação de integridade, rotulagem semântica, controle de qualidade, particionamento e versionamento de artefatos, e não um modelo preditivo isolado. O objetivo é estruturar uma cadeia auditável que conecte origem dos dados, validação de integridade, extração de rótulos, alinhamento matriz–metadados, controle de qualidade, particionamento antivazamento e geração de artefatos versionados. A proposta está alinhada aos princípios FAIR, aqui entendidos como dados Encontráveis, Acessíveis, Interoperáveis e Reutilizáveis [Wilkinson et al. 2016].

Metodologicamente, o *pipeline* foi aplicado a séries públicas do GEO (*Gene Expression Omnibus*) [Edgar et al. 2002], com separação entre coortes teciduais externas e linhagens líquidas. O fluxo registra proveniência, aplica validação de integridade, normaliza metadados textuais, define regras determinísticas de rotulagem, alinha matrizes e anotações, executa controle de qualidade e preserva resumos tabulares em JSON/CSV. Essa organização busca tornar explícitas decisões que frequentemente ficam dispersas em *scripts* ou planilhas intermediárias.

As principais contribuições organizam-se em três dimensões. A primeira é um *pipeline* reprodutível e auditável de engenharia de dados em coortes públicas de oncologia pulmonar. A segunda é a separação explícita de domínios biológicos e de políticas de particionamento anti-vazamento, evitando comparação indevida entre tarefas. A terceira é a apresentação de evidências verificáveis de execução em coortes externas, com consolidação de GSE43458 (110 amostras) e GSE32863 (116 amostras), totalizando 226 amostras com rótulos binários completos e zero amostras não classificadas nas coortes incluídas. Assim, a contribuição para Ciência de Dados para o Bem Social está em oferecer infraestrutura metodológica que pode apoiar grupos de pesquisa na produção de evidências mais comparáveis e reprodutíveis com dados públicos.

Além desta introdução, a Seção 2 apresenta trabalhos relacionados. A Seção 3 descreve os métodos e as fontes públicas utilizadas. A Seção 4 sintetiza o *pipeline* proposto. A Seção 5 apresenta os resultados da execução auditada. As Seções 6 e 7 discutem implicações, limitações e conclusões do estudo.

2. Trabalhos Relacionados

A literatura sobre aprendizado de máquina em saúde concentra-se, com razão, em arquiteturas preditivas e desempenho em coortes retrospectivas. Entretanto, revisões e diretrizes

convergem para um ponto menos visível na prática, mas determinante para a validade da evidência. Qualidade, governança e rastreabilidade dos dados precedem a escolha do algoritmo [Beam and Kohane 2018, Collins et al. 2015, Wolff et al. 2019]. Nesse sentido, este trabalho se aproxima menos da tradição de novos classificadores clínicos e mais de infraestruturas metodológicas que tornam pesquisas comparáveis, auditáveis e reutilizáveis em saúde pública.

Os princípios FAIR formalizam que dados científicos devem ser Encontráveis, Acessíveis, Interoperáveis e Reutilizáveis [Wilkinson et al. 2016]. Em biomedicina computacional, isso não se limita à disponibilização de arquivos brutos. Envolve manifestos de integridade, versionamento de *snapshots* e documentação de transformações. De modo complementar, TRIPOD e PROBAST reforçam que decisões de desenho dos dados, particionamento, seleção de variáveis e reporte transparente integram a própria evidência científica, especialmente quando modelos são avaliados em cenários diagnósticos ou prognósticos [Collins et al. 2015, Wolff et al. 2019]. Essas recomendações dialogam com boas práticas de aprendizado estatístico, nas quais transformações aprendidas com informação do teste podem inflar artificialmente o desempenho [Hastie et al. 2009, Steyerberg and Vergouwe 2014].

Repositórios públicos, como o GEO [Edgar et al. 2002], democratizam o acesso a dados biomédicos, mas introduzem desafios recorrentes de engenharia. Entre eles estão metadados textuais heterogêneos, identificadores diferentes entre matriz de expressão e anotação SOFT, *downloads* parciais, substituição silenciosa de arquivos e rotulagem clínica ambígua. Coortes teciduais consolidadas, como as do Atlas do Genoma do Câncer para adenocarcinoma pulmonar [Cancer Genome Atlas Research Network 2014], mostram o valor científico de dados públicos bem curados, mas também evidenciam que harmonização e rastreabilidade são camadas metodológicas indispensáveis quando fontes heterogêneas são combinadas.

Por fim, discussões sobre transparência algorítmica em decisões de alto impacto reforçam que auditabilidade deve envolver não apenas o modelo, mas também a cadeia que produz os dados analisados [Rudin 2019]. Neste artigo, a interpretabilidade recai primariamente sobre a cadeia de custódia dos dados, incluindo regras de rotulagem verificáveis, exclusão explícita de classes ambíguas, separação de domínios biológicos e preservação de artefatos versionados.

Em síntese, os trabalhos e diretrizes discutidos oferecem fundamentos para governança, reporte, avaliação de risco de viés, transparência e reutilização de dados biomédicos. Entretanto, eles não descrevem, de forma integrada, um fluxo operacional voltado à ingestão, validação de integridade, rotulagem determinística, alinhamento matriz–metadados, controle de qualidade, particionamento anti-vazamento e versionamento de artefatos em coortes públicas de oncologia pulmonar. O pipeline proposto oferece um protocolo reproduzível para que grupos distintos possam produzir, verificar e reutilizar evidências em oncologia pulmonar sem depender de bases proprietárias ou curadoria opaca.

3. Métodos

O *pipeline* foi projetado com quatro princípios. São eles o determinismo operacional, a separação de domínios biológicos, a reprodutibilidade antes de performance e a cadeia de

custódia dos artefatos. Cadeia de custódia, neste contexto, significa manter o histórico verificável de origem, processamento e versão dos arquivos. Cada etapa opera com entradas explícitas, regras fixas e saídas versionadas, seguindo recomendações de governança de dados reprodutíveis [Wilkinson et al. 2016, Steyerberg and Vergouwe 2014]. A linhagem líquida foi tratada em domínios separados, como TEP (*tumor-educated platelets*, ou plaquetas educadas por tumor) e cfRNA (*cell-free RNA*, ou RNA livre de células), para evitar reutilização indevida de rótulos, partições e métricas entre contextos biologicamente distintos.

Todos os insumos utilizados são secundários e públicos, obtidos de repositórios internacionais de expressão gênica e metadados associados, sem coleta primária de material biológico pelos autores. As séries foram recuperadas do GEO/NCBI [Edgar et al. 2002], via *accession* público, isto é, o identificador persistente de uma série GEO, e artefatos associados. Entre eles estão arquivos *series matrix*, ficheiros tabulares com metadados de amostra e, conforme a plataforma, valores de expressão, matrizes de contagem e metadados SOFT. SOFT (*Simple Omnibus Format in Text*) é o formato textual do GEO para metadados de séries, plataformas e amostras. A Tabela 1 resume as fontes consideradas no desenho do pipeline, seus estudos primários e o papel metodológico de cada domínio. Essa explicitação reforça encontrabilidade e citação responsável das bases originais, além de permitir que terceiros localizem os mesmos dados brutos independentemente da implementação computacional do fluxo.

Tabela 1. Fontes públicas e estudos de origem utilizados no pipeline.

Fonte	Estudo de origem	Modalidade	Papel no *pipeline*
GSE43458	Kabbout et al. [Kabbout et al. 2013]	tecido pulmonar, *microarray*	validação externa tecidual; LUAD vs. não tumoral
GSE32863	Selamat et al. [Selamat et al. 2012]	tecido pulmonar, *microarray*	validação externa tecidual; LUAD vs. adjacente não tumoral
GSE89843	Best et al. [Best et al. 2017]	TEP, *RNA-seq*	domínio líquido; rotulagem NSCLC vs. controle saudável
GSE216561	Wang et al. [Wang et al. 2024]	cfRNA plasma/EV, *RNA-seq*	domínio líquido; agrupamento por sujeito e fração biológica

A metodologia adotou um eixo conceitual em que decisões de modelagem e inferência só são consideradas válidas quando vinculadas a um ciclo explícito de qualidade e proveniência de dados. Na prática, isso implica manter rastreabilidade entre origem do dado, transformação aplicada e resultado reportado. Em termos operacionais, os artefatos de engenharia de dados, incluindo manifestos de integridade, tabelas de alinhamento, partições e saídas processadas, estruturam uma base consistente para etapas analíticas e de avaliação, reduzindo ambiguidades de execução e fortalecendo a reprodutibilidade.

O fluxo foi organizado em seis etapas sequenciais. Primeiro, ocorre a ingestão de arquivos brutos em fontes públicas. Em seguida, cada recurso passa por validação de integridade por SHA-256, algoritmo de *hash* usado para verificar se o conteúdo de um arquivo foi preservado, com registro de proveniência. A terceira etapa realiza *parsing*, isto é, leitura estruturada de campos textuais, e normalização de metadados para construção de rótulos. Depois, o *pipeline* executa alinhamento determinístico entre a matriz de contagens e os metadados, controle de qualidade tabular e preparação analítica. Por fim, os dados são particionados em treino e teste com estratégia anti-vazamento. Cada etapa gera artefatos intermediários persistidos para auditoria.

A ingestão de fontes públicas foi estruturada como processo reproduzível, com validação criptográfica por SHA-256 e registro de origem, caminho relativo, *hash* e contexto biológico. Essa camada evita dependência de *downloads ad hoc*, reduz risco de corrupção silenciosa em repositórios de dados biomédicos [Edgar et al. 2002] e permite reconstruir o estado exato da entrada usada em cada execução.

A extração semântica de rótulos parte de metadados textuais normalizados por regras determinísticas para identificar doença, coorte e atributos operacionais. A construção do alvo binário adota critério estrito. Amostras sem informação diagnóstica suficiente são excluídas da tarefa específica. Nas coortes teciduais externas, as classes tumor e não tumor foram derivadas de campos de metadados diagnósticos e descritivos do GEO, com mapeamento determinístico de termos compatíveis com tecido tumoral, tecido normal adjacente ou controle não tumoral. Registros sem correspondência inequívoca são mantidos fora da tarefa binária, evitando atribuição manual subjetiva e reduzindo risco de rótulos espúrios. No caso TEP, o alinhamento entre colunas da matriz e anotações SOFT (*Simple Omnibus Format in Text*, formato textual usado para organizar metadados no GEO) exige transformação determinística de identificadores, reduzindo risco de troca de rótulo, zeros artificiais e inclusão de amostras incompatíveis com o desfecho. Esse passo foi tratado como requisito metodológico crítico, pois erros de mapeamento comprometem todas as análises *downstream*, termo usado para indicar etapas posteriores do fluxo analítico.

O controle de qualidade (QC) é executado em leitura por *chunks*, ou blocos de linhas processados em partes, permitindo robustez operacional em arquivos extensos e cálculo incremental de indicadores mínimos: dimensões efetivas, volume de células numéricas, percentual de ausentes e estatísticas descritivas. Além dos indicadores globais, o *pipeline* verifica consistência de tipos, variação inesperada de colunas entre blocos e presença de padrões compatíveis com truncamento de arquivo. Apenas após essas verificações os dados seguem para preparação analítica. Após alinhamento e filtragem, a preparação numérica aplica transformações adequadas e seleciona variáveis de alta variância apenas no conjunto de treino, reduzindo vazamento de informação do teste e viés otimista na avaliação [Steyerberg and Vergouwe 2014, Hastie et al. 2009].

O particionamento contempla, quando aplicado, *split* estratificado, *split* por grupo e *split* estratificado por grupo. A configuração operacional registrada nos artefatos usa proporção 80% treino e 20% teste, com semente fixa 42, salvo quando a preservação de grupos impõe ajuste de cardinalidade. *Split* significa divisão do conjunto de dados em subconjuntos de treino e teste; a forma estratificada preserva a distribuição das classes, enquanto a forma por grupo mantém amostras relacionadas no mesmo subconjunto. Essa decisão reduz pseudo-replicação e otimismo artificial em cenários clínicos retrospectivos [Collins et al. 2015, Wolff et al. 2019]. Cada execução produz artefatos auditáveis de treino, teste e atribuição consolidada. A camada processada é congelada por versão temporal, com manifestos contendo origem, caminho relativo, *hash* e contexto biológico. Assim, toda métrica reportada deve ser reconduzível ao mesmo *snapshot*, isto é, ao mesmo retrato congelado do estado dos dados, à mesma estratégia de *split* e ao mesmo conjunto de regras de transformação.

A partir desses princípios e fontes públicas, o fluxo operacional foi organizado como uma cadeia auditável de etapas dependentes. A próxima seção apresenta o *pipeline* proposto, sua lógica sequencial e os artefatos gerados em cada etapa.

4. Pipeline Proposto

O fluxo foi organizado para que cada etapa dependa de uma saída verificável da etapa anterior, preservando uma trilha de auditoria composta por origem dos dados, *hash*, regras de rotulagem, controle de qualidade, particionamento e versão dos artefatos.

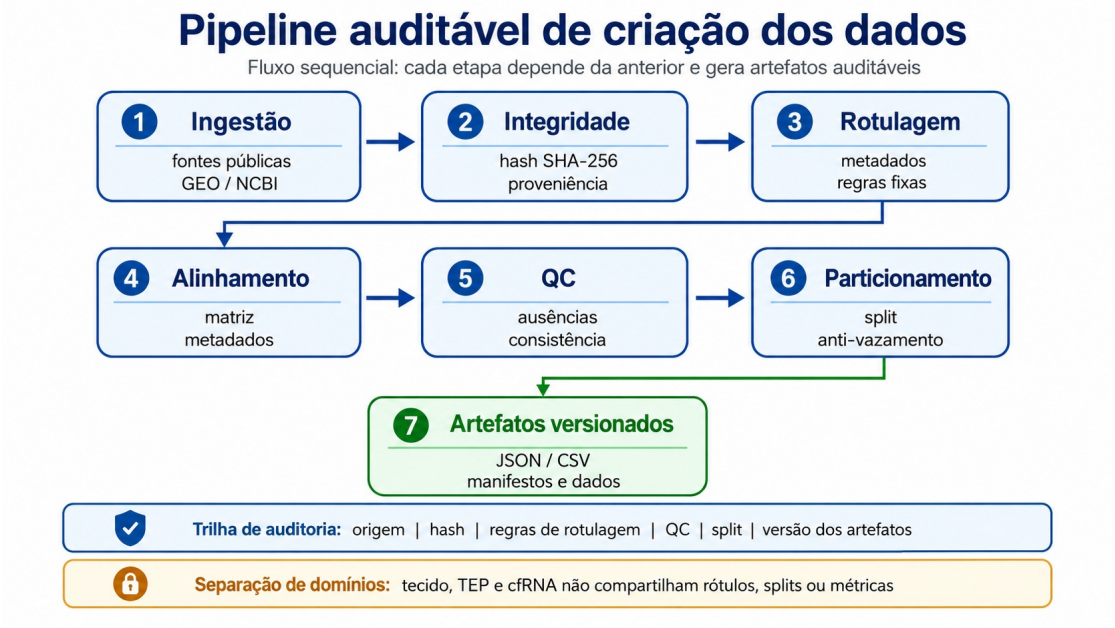

Figura 1. Fluxo operacional de criação e versionamento dos dados públicos biomédicos.

A cadeia operacional segue seis regras principais. Primeiro, cada arquivo bruto é associado ao repositório público de origem e a um registro de integridade. Segundo, metadados textuais são normalizados por regras fixas, sem rotulagem manual discricionária. Terceiro, a matriz e os metadados são alinhados antes de qualquer estatística descritiva ou modelagem. Quarto, o controle de qualidade registra dimensões efetivas, ausências e sinais de inconsistência estrutural. Quinto, o particionamento é tratado como artefato de dados, não como detalhe posterior de modelagem. Por fim, cada execução congela resumos JSON/CSV, metadados de versão e saídas processadas, permitindo que as contagens reportadas sejam reconduzidas à mesma execução auditada.

Essa organização evita que decisões críticas, como rotulagem, alinhamento matriz–metadados e particionamento, fiquem dispersas ou implícitas no código. Além disso, a separação entre domínios biológicos impede que tecido, TEP e cfRNA compartilhem rótulos, partições ou métricas de forma indevida. Cada etapa gera artefatos auditáveis, incluindo manifestos de integridade, tabelas de alinhamento, resumos de controle de qualidade, partições e saídas processadas. A proposta, portanto, deve ser lida como infraestrutura metodológica para criação de bases auditáveis, e não como resultado preditivo final.

5. Resultados da execução auditada

A validação empírica reportada nesta seção consiste na execução auditada do pipeline sobre coortes públicas de oncologia pulmonar, e não na avaliação final de classificadores

preditivos. Assim, a unidade de análise é o próprio processo de engenharia de dados, verificado por meio de proveniência, integridade criptográfica, regras de rotulagem, separação de domínios, contagens reproduzíveis e artefatos versionados.

A execução ocorreu em ambiente local Windows, com Python 3.14.3 e bibliotecas pandas, scikit-learn e joblib, nos dias 24 e 25 de maio de 2026. Os artefatos de reprodução incluem manifestos SHA-256, regras de rotulagem, resumos CSV/JSON, auditorias de separação de domínios e instruções de execução. Os dados brutos não são redistribuídos; sua reconstrução parte dos identificadores públicos do GEO/NCBI e dos hashes documentados.

Na execução auditada para coortes externas de tecido, duas bases públicas foram consolidadas com rótulos binários completos. A definição operacional de classe foi realizada a partir dos campos diagnósticos e descritores de amostra disponíveis nos metadados públicos, classificando como tumorais as amostras explicitamente associadas a tecido tumoral pulmonar ou adenocarcinoma pulmonar e como não tumorais aquelas descritas como tecido normal, controle ou adjacente não tumoral, excluindo registros ambíguos ou sem informação suficiente. Na normalização textual, padrões de não tumor foram avaliados antes de padrões tumorais, reduzindo falsos positivos por substring, como em expressões do tipo *non-tumor*. A coorte GSE43458 totalizou 110 amostras, sendo 80 tumorais e 30 não tumorais. A coorte GSE32863 totalizou 116 amostras, com 58 tumorais e 58 não tumorais. Em ambas, o artefato de sumarização registrou zero amostras não classificadas, indicando que, após a aplicação das regras de rotulagem usadas nessa etapa, todas as amostras incluídas na consolidação possuíam classe operacional definida.

A remoção das métricas preditivas da tabela é deliberada. Embora existam artefatos parciais com divisões de treino e teste para linhagens líquidas, esses valores pertencem a outro domínio biológico e não descrevem a criação das coortes teciduais externas reportadas aqui. Mantê-los junto das evidências de construção dos dados poderia sugerir comparação indevida entre tarefas, o que contrariaria o princípio de separação de domínios adotado no pipeline.

Tabela 2. Evidências auditáveis da execução do pipeline.

Dimensão auditada	Artefato/evidência	Resultado
Integridade	Manifestos SHA-256	origem, caminho e hash registrados
Consolidação externa	GSE43458	110: 80 tumorais / 30 não tumorais
Consolidação externa	GSE32863	116: 58 tumorais / 58 não tumorais
Rotulagem	Regras determinísticas	0 não classificadas
Separação de domínios	Auditoria tecido/TEP/cfRNA	sem mistura de métricas
Reprodutibilidade	CSV/JSON e instruções	artefatos versionados

Os resultados da Tabela 2 evidenciam, no escopo da execução auditada, que o pipeline produz contagens verificáveis de amostras classificadas, regras explícitas de exclusão e rastreabilidade até artefatos versionados de execução. A ausência de amostras não classificadas nas coortes teciduais auditadas indica que as heurísticas textuais, após refinamento, cobriram integralmente o conjunto incluído na consolidação. Ao mesmo tempo, a manutenção separada de linhagens líquidas explicita que total de amostras, elegibilidade binária e desempenho preditivo pertencem a camadas distintas do ciclo analítico.

Do ponto de vista metodológico, os ganhos aparecem na rastreabilidade e na de-

fensabilidade da base resultante. A origem das contagens, a regra de formação dos rótulos, a exclusão de classes ambíguas, o alinhamento matriz–metadados e a preservação dos artefatos de versão deixam de ser decisões implícitas. Com isso, a interpretação dos dados passa a depender de uma cadeia verificável de criação, e não apenas de uma planilha final ou de uma métrica de desempenho.

6. Discussão

A principal implicação do estudo é que desempenho preditivo, sozinho, não sustenta credibilidade translacional. Em ambiente clínico, engenharia de dados não deve ser tratada como etapa acessória. Ela define, em grande medida, a validade interna e externa do que será reportado [Beam and Kohane 2018, Collins et al. 2015]. Em outras palavras, um modelo pode apresentar boa métrica e ainda assim falhar em confiabilidade se a cadeia de dados não estiver sob controle.

Sob a perspectiva de Ciência de Dados para o Bem Social, o pipeline proposto é relevante porque reprodutibilidade e auditabilidade não são apenas requisitos técnicos. Elas afetam quem consegue produzir, verificar e reutilizar evidências em saúde pública. Pipelines documentados com dados públicos podem diminuir dependência de bases proprietárias, favorecer comparação entre grupos e reduzir o risco de que decisões de política científica ou clínica sejam influenciadas por resultados inflados por vazamento ou por curadoria pouco transparente.

Essa preocupação é coerente com a literatura de validação preditiva, que associa o desenvolvimento de bons modelos à validação cuidadosa e ao desenho adequado dos dados [Steyerberg and Vergouwe 2014]. De modo complementar, PROBAST, ferramenta de avaliação de risco de viés em modelos preditivos, explicita que vazamento, seleção inadequada de amostras e decisões pós-hoc comprometem a interpretação clínica do desempenho [Wolff et al. 2019].

No presente trabalho, as ameaças residuais permanecem claras. Na validade de construto, rótulos derivados de metadados públicos podem carregar ambiguidades da anotação original. Na validade interna, ainda há risco de vazamento quando transformações estatísticas são aprendidas fora do treino. Na validade externa, o *domain shift*, isto é, a mudança de distribuição entre coortes públicas e cenário assistencial real, limita generalização.

Há também limitações sociais e de equidade. Coortes públicas podem sub-representar populações, serviços e territórios mais vulneráveis; portanto, a disponibilidade pública do dado não garante representatividade. O papel do pipeline é tornar essas restrições mais visíveis e auditáveis, não eliminá-las. Estudos futuros devem documentar origem populacional, critérios de inclusão, distribuição demográfica quando disponível e efeitos de validação externa antes de qualquer uso assistencial.

Do ponto de vista operacional, mudanças em fontes públicas e substituição de arquivos reforçam que ingestão idempotente com verificação de *hash* deve ser requisito permanente, não medida eventual. Por isso, as boas práticas propostas aproximam o fluxo tanto de FAIR quanto de recomendações de reporte transparente para modelos preditivos [Wilkinson et al. 2016, Collins et al. 2015, Wolff et al. 2019]. Essas práticas incluem separação estrita de domínios biológicos, *parsing* versionado, *split* compatível com estrutura de grupo e registro sistemático de proveniência.

7. Conclusão

Este trabalho corrobora a necessidade de tratar a criação de *datasets* derivados de coortes públicas, auditáveis e preparados para pesquisa clínico-computacional como uma etapa crítica da pesquisa em saúde pública. O principal legado é um processo reproduzível que integra verificação de integridade por SHA-256, semântica de rótulos, controle de qualidade escalável, particionamento defensivo e governança de versão.

Em termos práticos, o *pipeline* transforma dados públicos heterogêneos em base analítica mais confiável para pesquisa em aprendizado de máquina clínico, com separação explícita de domínios e trilha de proveniência ponta a ponta. O *pipeline* pode contribuir para reduzir barreiras técnicas para grupos de pesquisa, favorecer reuso responsável de dados abertos e fortalecer a transparência necessária a aplicações de Ciência de Dados para o Bem Social.

A integração entre engenharia de dados, modelagem e operação fortalece a coerência científica do *pipeline*, pois conecta decisões metodológicas e resultados com um mesmo contrato de reprodutibilidade. Ao explicitar limitações de domínio, risco de vazamento e proveniência, o trabalho oferece uma base mais defensável para estudos futuros de diagnóstico precoce, vigilância epidemiológica e avaliação responsável de modelos em oncologia pulmonar.

Como limitação, a execução auditada avalia a robustez do processo de engenharia de dados e não substitui validação clínica prospectiva, comparação sistemática com outros pipelines ou análise completa de equidade populacional. O pipeline também depende da qualidade e completude dos metadados públicos, e sua auditoria plena requer o pacote suplementar com manifestos e resumos de execução. Como próximos passos, recomenda-se ampliar validações externas em novos domínios, preservando a separação entre validação do pipeline de dados e avaliação de desempenho clínico de modelos preditivos.

8. Declarações

Uso de inteligência artificial: ferramentas de IA generativa foram utilizadas exclusivamente para apoio editorial, revisão linguística, padronização terminológica e melhoria de clareza textual. Todas as análises, interpretações, referências, resultados e conclusões foram verificadas e assumidas integralmente pelos autores humanos.

Disponibilidade de dados e artefatos: os insumos brutos são públicos e obtidos do NCBI GEO (GSE43458, GSE32863, GSE89843 e GSE216561), por meio dos respectivos identificadores GEO e da consulta pública `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43458`. Os artefatos derivados da execução auditada incluem manifestos de integridade, resumos CSV/JSON, regras de rotulagem e instruções de reprodução.

Ética: o estudo utiliza dados públicos desidentificados e secundários. O *pipeline* não substitui julgamento clínico nem validação prospectiva.

Referências

Beam, A. L. and Kohane, I. S. (2018). "Big data and machine learning in health care". *JAMA*, 319(13):1317–1318.

Best, M. G., Sol, N., In 't Veld, S. G. J. G., Vancura, A., Muller, M., Niemeijer, A.-L. N. et al. (2017). "Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets". *Cancer Cell*, 32(2):238–252.e9.

Kabbout, M., Garcia, M. M., Fujimoto, J., Liu, D. D., Woods, D., Chow, C.-W. et al. (2013). "ETS2 mediated tumor suppressive function and MET oncogene inhibition in human non-small cell lung cancer". *Clinical Cancer Research*, 19(13):3383–3395.

Selamat, S. A., Chung, B. S., Girard, L., Zhang, W., Zhang, Y., Campan, M. et al. (2012). "Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression". *Genome Research*, 22(7):1197–1211.

Wang, H., Zhao, W., Zhang, Y., Li, M. and Wang, P. (2024). "Depletion-assisted multiplexed cell-free RNA sequencing reveals distinct human and microbial signatures in plasma versus extracellular vesicles". *Clinical and Translational Medicine*, 14(7):e1760.

Cancer Genome Atlas Research Network (2014). "Comprehensive molecular profiling of lung adenocarcinoma". *Nature*, 511(7511):543–550.

Collins, G. S., Reitsma, J. B., Altman, D. G. and Moons, K. G. M. (2015). "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement". *Annals of Internal Medicine*, 162(1):55–63.

Edgar, R., Domrachev, M. and Lash, A. E. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". *Nucleic Acids Research*, 30(1):207–210.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, 2nd edition.

Siegel, R. L., Giaquinto, A. N. and Jemal, A. (2024). "Cancer statistics, 2024". *CA: A Cancer Journal for Clinicians*, 74(1):12–49.

Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". *Nature Machine Intelligence*, 1:206–215.

Steyerberg, E. W. and Vergouwe, Y. (2014). "Towards better clinical prediction models: seven steps for development and an ABCD for validation". *European Heart Journal*, 35(29):1925–1931.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*, 3:160018.

World Health Organization (2024). "Global cancer burden growing, amidst mounting need for services". WHO news release, 1 February 2024.

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S. et al. (2019). "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies". *Annals of Internal Medicine*, 170(1):51–58.