Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1C45530AD

Title: RadIA: Desenvolvimento de um modelo de inteligência artificial para previsão de planejamento em Radioterapia

Creator: Nathalia Aidar - **ORCID:** 0009-0005-2122-5705

Affiliation: Universidade de São Paulo (www5.usp.br)

Funder: Institutos Nacionais de Ciência e Tecnologia (inct.cnpq.br)

Funding opportunity number: INCT INTERAS - PROCESSO CNPq 406761/2022-1

Grant: INCT INTERAS - PROCESSO CNPq 406761/2022-1

Template: Template USP - Baseado no DCC

Project abstract:

Este Plano de Gestão de Dados (PGD) descreve os procedimentos para o gerenciamento do ciclo de vida dos dados utilizados no estudo sobre a aplicação de algoritmos de aprendizado de máquina (ML) para prever os resultados do controle de qualidade específico do paciente (PSQA) em radioterapia (IMRT e VMAT). O projeto analisa uma coorte de planos de radioterapia clínicos e anonimizados, abrangendo diversos locais de tratamento (próstata, cabeça e pescoço, pelve, SNC e SRS). A partir de cada plano DICOM-RT, são extraídas 60 features quantitativas que englobam a dinâmica do colimador multi-lâminas (MLC), características da distribuição de dose e índices de complexidade. Adicionalmente, para lidar com o desbalanceamento de classes, são gerados dados sintéticos, incluindo um conjunto de 66 planos "perturbados" e amostras criadas pela técnica SMOTE. Este PGD detalha como esses dados complexos e heterogêneos serão coletados, documentados, armazenados, protegidos e preparados para compartilhamento, garantindo a reprodutibilidade e o valor a longo prazo da pesquisa.

Start date: 08-01-2024

End date: 08-01-2026

Last modified: 09-18-2025

RadIA: Desenvolvimento de um modelo de inteligência artificial para previsão de planejamento em Radioterapia - Coleta de Dados

Detalhes dos dados coletados ou criados

Que dados serão coletados ou criados?

Dados Coletados (Primários): A base de dados primária consiste em uma coorte de planos de radioterapia clínicos e anonimizados. Esses planos abrangem uma variedade de locais de tratamento, incluindo próstata, cabeça e pescoço, pelve, sistema nervoso central (SNC) e radiocirurgia estereotáxica (SRS). Os dados brutos são provenientes de arquivos no padrão DICOM-RT (RTPLAN, RTDOSE, RTIMAGE).

Dados Criados/Gerados (Secundários): A partir dos dados primários, os seguintes dados serão criados ao longo do projeto:

Features Quantitativas: Para cada plano, serão extraídas 60 métricas quantitativas. Essas métricas incluem características da dinâmica do colimador multi-lâminas (MLC), da distribuição de dose e índices de complexidade como Modulation Complexity Score (MCS), Average Leaf Speed (ALS) e Leaf Sequence Variability (LSV).

Dataset Estruturado: As features extraídas serão organizadas em um banco de dados estruturado utilizando a biblioteca pandas do Python para manipulação e análise.
Dados Sintéticos para Balanceamento: Para tratar o desbalanceamento de classes (sucesso/falha em QA), serão gerados dois tipos de dados sintéticos:

Um conjunto de 66 planos "perturbados" que simulam desvios realistas na entrega do tratamento.
Amostras sintéticas da classe minoritária (falha em QA) serão geradas utilizando a técnica SMOTE (Synthetic Minority Oversampling Technique)

Como os dados serão coletados ou criados?

Coleta e Extração de Dados Primários: O processo de extração dos dados a partir dos arquivos DICOM-RT foi realizado através de um pipeline modular desenvolvido em Python.

Ferramentas: Foram utilizadas bibliotecas especializadas como pydicom para a análise e extração de informações dos arquivos DICOM, e pandas para organizar os dados extraídos em um formato estruturado.
Processo Modular: Scripts dedicados foram criados para processar cada modalidade de arquivo DICOM de forma independente:

RTPLAN: Um parser foi responsável por extrair configurações dos feixes de tratamento, tempos de entrega e doses prescritas.

RTDOSE: O parser de arquivos RTDOSE converteu os dados brutos de voxels em matrizes tridimensionais, permitindo o cálculo de parâmetros dosimétricos agregados.

RTIMAGE: O módulo RTIMAGE recuperou metadados das imagens portais, como os tempos de aquisição e valores de intensidade dos pixels.

Validação de Dados: Um módulo de validação foi integrado ao pipeline para verificar a integridade dos arquivos de entrada, excluindo conjuntos de dados corrompidos ou incompletos. Scripts compararam metadados extraídos com os valores esperados dos protocolos clínicos para identificar inconsistências.

2. Criação de Dados Sintéticos e Derivados: Para aumentar a robustez do modelo de machine learning, dados sintéticos foram gerados.

Simulação de Falhas: Foi criado um conjunto de 66 planos de tratamento sintéticos para simular falhas de QA. Isso foi feito aplicando perturbações controladas aos dados originais, como alteração das trajetórias do MLC, tempos de feixe e distribuições de dose.
Balanceamento de Classes: Para mitigar o desbalanceamento entre os casos de "sucesso" e "falha", foi utilizada a técnica SMOTE (Synthetic Minority Oversampling Technique). O SMOTE gera novas amostras da classe minoritária (falhas) através da interpolação entre as observações existentes no espaço de features.

3. Otimização e Reprodutibilidade:

Otimização de Desempenho: O pipeline foi otimizado para processamento em lote e computação paralela através da biblioteca multiprocessing do Python, reduzindo o tempo de execução.

Rastreabilidade: Uma rotina de logging foi incorporada para registrar o progresso, erros de execução e informações de diagnóstico, aumentando a rastreabilidade e a reprodutibilidade de todo o processo.

RadIA: Desenvolvimento de um modelo de inteligência artificial para previsão de planejamento em Radioterapia - Documentação e Metadados

Informações sobre Documentação e Metadados

Que documentação e metadados irão acompanhar os dados

Metadados Incorporados (Embedded Metadata):

Metadados DICOM-RT: Os dados primários, em formato DICOM-RT, contêm metadados ricos e padronizados incorporados aos próprios arquivos. Isso inclui um vasto conjunto de informações, como:

Configurações do feixe de tratamento, tempos de entrega e doses prescritas (extraídos dos arquivos RTPLAN).
Parâmetros dosimétricos agregados, como dose média e máxima (computados a partir dos arquivos RTDOSE).
Timestamps de aquisição e valores de intensidade de pixel (recuperados dos arquivos RTIMAGE).
Dose prescrita, dose máxima, número de frações e Unidades Monitoras (MU) totais.
2. Documentação e Metadados Descritivos: Será produzida uma documentação detalhada para descrever o contexto, a estrutura e o processo de geração dos dados.

Dicionário de Dados das Features: Uma documentação completa descrevendo cada uma das 60 features quantitativas extraídas dos planos. Essa documentação explicará o que cada feature representa. Exemplos de features documentadas incluem o Modulation Complexity Score (MCS), Average Leaf Speed (ALS) , e Fator de Irregularidade.

Documentação do Processo (Logs): O pipeline computacional inclui uma rotina de logging que captura o progresso do processamento, erros de execução e informações de diagnóstico. Esses logs servem como metadados do processo, aumentando a rastreabilidade e a reprodutibilidade.

Código-Fonte Comentado: Os scripts em Python utilizados para a extração, processamento e análise dos dados serão comentados para explicar a lógica da implementação e o fluxo de trabalho.

RadIA: Desenvolvimento de um modelo de inteligência artificial para previsão de planejamento em Radioterapia - Conformidade Ética e Legal

Informações sobre Conformidade Ética e Legal

Como serão tratadas as questões éticas e legais?

estudo utiliza uma coorte planos de radioterapia clínicos que foram anonimizados antes de serem incluídos na análise. Este procedimento garante a proteção da confidencialidade e da privacidade dos pacientes, uma vez que todas as informações de identificação pessoal foram removidas dos arquivos DICOM-RT, em conformidade com as boas práticas de pesquisa clínica. Não foi necessário o uso de comite de ética já que nenhum dado sensível de paciente é utilizado nessa pesquisa.

Como serão tratadas as questões de direito e propriedade intelectual do autor?

Direito Autoral da Publicação: A propriedade intelectual e os direitos autorais do trabalho científico (o artigo) são atribuídos aos autores listados na publicação.

Propriedade Intelectual do Projeto (Dados e Modelo): A gestão dos direitos sobre os dados brutos, os dados gerados e o modelo de machine learning é fundamentada na parceria formal entre as instituições envolvidas.

Conflito de Interesses: Conforme informado, as partes envolvidas na parceria declaram não ter conflito de interesses, o que reforça a natureza estritamente científica e colaborativa do trabalho, visando o avanço do conhecimento na área.

RadIA: Desenvolvimento de um modelo de inteligência artificial para previsão de planejamento em Radioterapia - Armazenamento, Backup, Responsabilidade e Recursos

Informações sobre Armazenamento e Backup

Como os dados serão armazenados e como serão realizadas as cópias de segurança durante a pesquisa?

Armazenamento Durante a Pesquisa (Ativo):

Armazenamento Primário: O conjunto de dados principal, incluindo os arquivos DICOM-RT anonimizados e o banco de dados de features, será mantido em um servidor de armazenamento seguro (NAS - Network Attached Storage) ou em um repositório institucional provido pelo IPEN. Este servidor terá controle de acesso restrito, garantindo que apenas os pesquisadores diretamente envolvidos no projeto possam acessar os dados.
Armazenamento Local (de Trabalho): Cópias de trabalho dos dados e dos scripts poderão ser mantidas nos computadores locais dos pesquisadores para fins de processamento e análise. Todas as máquinas locais utilizadas terão o disco rígido criptografado para proteger os dados em caso de perda ou roubo do equipamento.

Cópias de Segurança (Backup):

Backup Regular Automatizado: Serão configurados backups automáticos e incrementais do servidor de armazenamento primário. Diariamente, as alterações serão salvas no mesmo servidor, e semanalmente, um backup completo será realizado.
Backup Externo (On-site): Uma cópia de segurança completa do projeto será mantida em um disco rígido externo criptografado, guardado em um local físico seguro e diferente do local do servidor principal (e.g., em um cofre no laboratório). Esta cópia será atualizada mensalmente.
Backup Remoto (Off-site): Uma terceira cópia de segurança será armazenada em um serviço de nuvem institucional (como Google Drive ou OneDrive, se oferecido pela USP/IPEN com garantias de segurança) ou em um servidor localizado em outra instalação física. Isso protege os dados contra incidentes locais, como incêndios ou falhas de energia generalizadas.

Como serão tratadas as questões de acesso e segurança?

Controle de Acesso Lógico:

Acesso Baseado em Função (Role-Based Access): O acesso aos dados brutos e ao banco de dados consolidado será estritamente controlado. Será implementado um sistema de permissões em que apenas os pesquisadores principais e a equipe diretamente envolvida na análise de dados terão acesso de leitura e escrita. Colaboradores ou estudantes terão acesso apenas a subconjuntos de dados anonimizados e necessários para suas tarefas específicas.
Autenticação Forte: O acesso ao servidor de armazenamento de dados exigirá autenticação via credenciais institucionais (login e senha), e, se a infraestrutura permitir, autenticação de dois fatores (2FA) será ativada para uma camada extra de segurança.

Segurança dos Dados:

Anonimização na Fonte: Conforme descrito no artigo, a medida de segurança mais importante é que todos os planos clínicos foram completamente anonimizados na origem. Este processo removeu todas as informações de identificação pessoal (PHI - Protected Health Information), sendo o passo fundamental para proteger a privacidade dos pacientes.
Criptografia:

Em Repouso (At Rest): Os dados armazenados no servidor principal e nos dispositivos de backup (discos externos e armazenamento em nuvem) serão protegidos com criptografia. Isso garante que, mesmo em caso de acesso físico não autorizado ao hardware, os dados permaneçam ilegíveis.
Em Trânsito (In Transit): Toda a transferência de dados entre os computadores locais dos pesquisadores e o servidor central será feita através de conexões seguras e criptografadas, como SFTP (Secure File Transfer Protocol) ou VPN (Virtual Private Network) institucional.

Informações sobre Responsabilidade e Recursos

Quem será responsável pelo gerenciamento dos dados?

Pesquisadora Principal (Principal Investigator - PI): A responsabilidade final pela gestão dos dados recai sobre a pesquisadora principal do projeto, Nathália L. A. Alves, e seu orientador/supervisor, Mário Olímpio de Menezes. Eles são responsáveis por garantir que o Plano de Gestão de Dados seja implementado, seguido e atualizado conforme necessário, e que todas as diretrizes éticas e de segurança sejam cumpridas.
Equipe de Pesquisa: A equipe de pesquisa do IPEN, incluindo os co-autores do trabalho, será responsável pelo gerenciamento diário dos dados. Isso inclui tarefas como a execução do pipeline de extração de dados, a organização dos arquivos, a realização de backups e a manutenção da documentação (como o dicionário de dados).
Parceiros Institucionais (RTCON e Hospital IBCC): Os parceiros institucionais são co-responsáveis pela gestão e segurança dos dados em sua origem. O Hospital IBCC é responsável por garantir que o processo de coleta e anonimização inicial dos dados dos pacientes siga as normas éticas e legais. A RTCON Solutions, como parceira tecnológica, compartilha a responsabilidade pela integridade e pelo correto manuseio dos dados dentro do escopo da colaboração técnica.

Que recursos serão necessários para manter esse plano?

Recursos de Software:

Software de Análise: O principal recurso de software é o ambiente de programação Python (versão 3.x ou superior) e suas bibliotecas científicas, que são de código aberto e não têm custo de licença. As bibliotecas essenciais, já utilizadas no projeto, incluem:

pydicom: para leitura e manipulação de arquivos DICOM.
pandas: para estruturação e manipulação do banco de dados de features.
scikit-learn: para a implementação do modelo Random Forest e da técnica SMOTE.
multiprocessing: para otimização e processamento paralelo.
Software de Controle de Versão: Será utilizado o Git para o controle de versão dos scripts de análise, garantindo a rastreabilidade das alterações e facilitando a colaboração.
Software de Backup: Utilização de softwares nativos ou de mercado para automatizar as rotinas de backup no servidor e nos dispositivos externos.
2. Recursos de Hardware:

Armazenamento Seguro: Acesso a um servidor de armazenamento de dados (NAS) ou a um repositório institucional do IPEN com capacidade inicial de 2-3 Terabytes, para acomodar os dados brutos, os dados gerados e as cópias de segurança.
Dispositivos de Backup: Aquisição de pelo menos um disco rígido externo com capacidade de 4TB para o backup on-site, e alocação de cota em um serviço de armazenamento em nuvem institucional para o backup off-site.
Recursos Computacionais: Acesso contínuo a estações de trabalho ou a um servidor com capacidade de processamento (CPU multi-core) e memória RAM (mínimo 16-32 GB) adequadas para executar o pipeline de extração de features e o treinamento dos modelos de machine learning, que são tarefas computacionalmente

intensivas.
3. Recursos de Pessoal e Tempo:

Pessoal Especializado: A execução do plano depende da disponibilidade contínua da equipe de pesquisa, que possui a expertise necessária em Física das Radiações, programação em Python e técnicas de machine learning.
Tempo Dedicado: Será alocado tempo específico dos pesquisadores para as atividades de gestão de dados, que incluem:

Curadoria e organização dos dados.
Manutenção e atualização da documentação (dicionário de dados, README).
Execução e verificação das rotinas de backup.
Preparação dos dados para preservação e compartilhamento ao final do projeto.

RadIA: Desenvolvimento de um modelo de inteligência artificial para previsão de planejamento em Radioterapia - Seleção, Preservação e Compartilhamento

Informações sobre Seleção e Preservação dos dados

Que dados são de longo prazo e precisarão ser mantidos, compartilhados e/ou preservados?

O Dataset Final de Features Quantitativas (CSV):

O que é: O arquivo de dados estruturado (em formato .csv) contendo as 60 métricas quantitativas extraídas para cada um dos planos clínicos, juntamente com o resultado correspondente do PSQA (passa/falha).
Por que preservar/compartilhar: Este é o ativo de dados mais valioso e reutilizável do projeto. Ele é anonimizado, relativamente pequeno em tamanho, fácil de usar e contém toda a informação necessária para que outros pesquisadores possam treinar seus próprios modelos, comparar resultados ou realizar novas análises estatísticas.
O Código-Fonte Completo (Scripts Python):

O que é: Todos os scripts em Python desenvolvidos para o pipeline de extração de features, geração de dados sintéticos (SMOTE), treinamento e validação do modelo Random Forest.
Por que preservar/compartilhar: A preservação do código é essencial para a reprodutibilidade. Com acesso a este código e ao dataset de features, qualquer pesquisador pode replicar exatamente a sua análise, validar seus resultados e adaptar a metodologia para outros conjuntos de dados.
A Documentação Associada:

O que é: O Dicionário de Dados que descreve cada uma das 60 features, e o arquivo README que explica como executar os scripts e a estrutura do projeto.
Por que preservar/compartilhar: Sem documentação, os dados e o código perdem o contexto e se tornam inúteis. Essa documentação é fundamental para que outros possam entender e reutilizar seu trabalho corretamente.
O Modelo de Machine Learning Treinado:

O que é: O arquivo final do modelo Random Forest treinado e otimizado.
Por que preservar/compartilhar: Compartilhar o modelo permite que outros o utilizem diretamente para fazer previsões em novos dados, sem a necessidade de retreiná-lo.

Qual o plano para preservação de longo prazo para o conjunto de dados?

Seleção de um Repositório de Dados Confiável:

Ao final do projeto, o conjunto de dados de longo prazo (Dataset de Features em CSV, código-fonte em Python e documentação associada) será depositado em um repositório de dados de pesquisa reconhecido. As opções prioritárias são:

Repositório Institucional: IPEN ou a USP em seus respectivos repositórios de dados de pesquisa (e.g., "Repositório de Dados da USP") que ofereça garantias de preservação a longo prazo.
Repositórios Públicos: podem ser utilizados repositórios públicos multidisciplinares de renome, como o Zenodo (https://zenodo.org/) ou o Figshare (https://figshare.com/). Esses repositórios são projetados para preservação digital e são amplamente utilizados pela comunidade científica.

Preparação dos Dados para Depósito:

Formatos de Arquivo Abertos: Todos os arquivos serão salvos em formatos abertos e não-proprietários para maximizar a acessibilidade futura (e.g., .csv para dados tabulares, .txt ou .md para documentação, .py para código).

Pacote de Depósito Completo: Os dados, códigos e documentação serão agrupados em um único "pacote de depósito". Este pacote conterá um arquivo README.txt na raiz, explicando o conteúdo de cada arquivo e como eles se relacionam, garantindo que o contexto da pesquisa seja preservado junto com os dados.
Atribuição de um Identificador Persistente (PID):

No momento do depósito no repositório escolhido, será gerado um Identificador de Objeto Digital (DOI - Digital Object Identifier) para o conjunto de dados. O DOI garante que o dataset tenha um link de citação permanente e estável, facilitando sua localização, citação e rastreamento de impacto ao longo do tempo.
Período de Retenção:

O conjunto de dados depositado será mantido e estará acessível por um período mínimo de 10 anos após a publicação do artigo final, em conformidade com as políticas de agências de fomento e boas práticas científicas.

Licença de Uso:

Aos dados será atribuída uma licença de uso aberta, como a Creative Commons CC-BY 4.0, que permite o reuso dos dados por outros pesquisadores, desde que a devida atribuição (citação) ao trabalho original seja feita.

Informações sobre Compartilhamento dos dados

Como os dados serão compartilhados?

Compartilhamento Público (Via Repositório de Dados): Os seguintes ativos de pesquisa serão compartilhados publicamente para garantir a máxima transparência e reprodutibilidade:

O que será compartilhado: O conjunto de dados final de features (em formato .csv), o código-fonte completo e comentado (scripts Python) e a documentação associada (dicionário de dados e arquivo README).

Como será compartilhado: Este pacote de dados será depositado em um repositório de dados confiável (como o Zenodo ou Figshare), onde receberá um Identificador de Objeto Digital (DOI) para garantir uma citação permanente e fácil localização.
Quando será compartilhado: Os dados serão tornados públicos no momento da publicação do artigo científico associado.
Licença de Uso: Os dados serão compartilhados sob uma licença aberta, como a Creative Commons Attribution 4.0 (CC-BY 4.0), que permite o reuso irrestrito para fins de pesquisa, desde que o trabalho original seja devidamente creditado.

Acesso Controlado (Dados Sensíveis):

O que não será compartilhado publicamente: Os arquivos DICOM-RT brutos e anonimizados.
Por que o acesso será restrito: Devido ao seu grande volume e por serem derivados diretamente de dados clínicos, mesmo que anonimizados, o compartilhamento público irrestrito não é apropriado.
Como solicitar acesso: O acesso a este conjunto de dados brutos poderá ser concedido a pesquisadores qualificados mediante solicitação direta ao pesquisador principal. O compartilhamento dependerá da assinatura de um Acordo de Transferência de Dados (Data Transfer Agreement - DTA) entre as instituições envolvidas, especificando o propósito do uso e garantindo a conformidade com as políticas de segurança e ética.

É necessária alguma restrição no compartilhamento dos dados?

Restrição para os Dados Brutos Anonimizados (Arquivos DICOM-RT):

Tipo de Restrição: Acesso controlado e restrito.
Justificativa: Apesar de estarem anonimizados, estes dados são derivados diretamente de registros clínicos de pacientes. Há um risco residual e uma responsabilidade ética associada à sua distribuição. Além disso, os acordos de parceria com as instituições fornecedoras dos dados (como o Hospital IBCC) podem impor limitações ao seu compartilhamento irrestrito.
Implementação: Estes dados não serão disponibilizados publicamente. O acesso só será concedido a pesquisadores qualificados através de um Acordo de Transferência de Dados (Data Transfer Agreement - DTA), que formaliza o propósito do uso e as obrigações do solicitante.
Restrição para os Dados Processados e Públicos (Dataset de Features e Código):

Tipo de Restrição: Requisito de atribuição (citação).
Justificativa: O objetivo é compartilhar estes dados da forma mais aberta possível para promover a ciência, mas é imperativo que o trabalho original seja reconhecido e creditado.
Implementação: Estes dados serão compartilhados publicamente sob uma licença Creative Commons Atribuição 4.0 Internacional (CC-BY 4.0). Esta licença permite que qualquer pessoa copie, distribua e crie trabalhos derivados a partir dos dados, para qualquer fim (inclusive comercial), desde que dê o crédito apropriado (faça a citação) aos autores originais

---

Planned Research Outputs

Text - "Produtos e Resultados da Pesquisa"

Uma publicação científica revisada por pares: O principal resultado será um artigo de pesquisa detalhando a metodologia, os resultados e as conclusões do estudo.
Um conjunto de dados curado e reutilizável: Inclui o conjunto de dados final e anonimizado em formato .csv, contendo as 60 features quantitativas extraídas e os resultados de QA correspondentes para os planos clínicos analisados. Este é um ativo valioso para a comunidade de pesquisa.
Um modelo de machine learning treinado: O classificador Random Forest final e otimizado, capaz de prever os resultados do controle de qualidade específico do paciente (PSQA).
Software e código-fonte: O conjunto completo de scripts Python desenvolvidos para o pipeline de extração de dados, engenharia de features, implementação do SMOTE e treinamento do modelo, garantindo a reprodutibilidade da pesquisa.
Documentação de suporte: Inclui o próprio Plano de Gestão de Dados, um arquivo README.txt e um dicionário de dados abrangente que descreve cada uma das 60 features, garantindo que os dados e o código sejam compreensíveis e reutilizáveis por outros.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Produtos e Resultados da Pesquisa	Text	2025-09-17	Open	Washington University Research Data AnVIL Apollo Australian Breast Cancer Tissue Bank		Creative Commons Attribution 4.0 International	FAIR Data Principles cancer Data Standards Registry and Repository (caDSR)	No	No