Capítulo IV

Proposta de governança de dados para publicações científicas e relatórios técnicos do Projeto Plataforma Clínica Global para a Covid-19 no Brasil

Proposta de governança de dados para publicações científicas e relatórios técnicos do Projeto Plataforma Clínica Global para a Covid-19 no Brasil

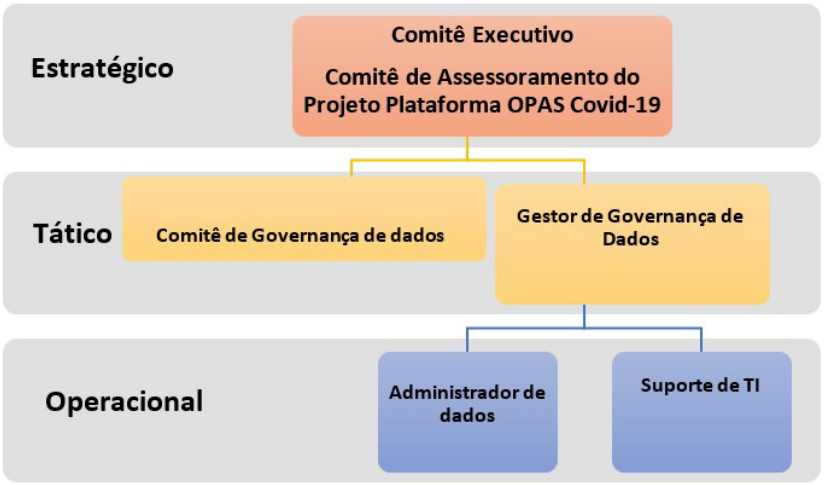
Ricardo Kuchenbecker, Gabriel Muller, Rafael Moraes, Leonardo Nunes Alegre, Natália Del Angelo Aredes, Rosane de Mendonça Gomes, Fernando Anschau e Eduardo Barbosa Coelho

A partir do projeto Plataforma Global de Dados Clínicos Covid-19 da Organização Mundial da Saúde, o Escritório da Organização Pan-Americana da Saúde no Brasil plantou o desafio de reunir um grupo de hospitais da Empresa Brasileira de Serviços Hospitalares (Ebserh), Hospital de Clínicas de Porto Alegre (HCPA), Grupo Hospitalar Conceição (GHC), Complexo Hospitalar do Trabalhador, Hospital da Criança de Brasília, Instituto de Saúde e Gestão Hospitalar (ISGH), e Hospital Getúlio Vargas (HGV SES-PI) para que fosse possível organizar mecanismos de processamento de dados assistenciais de pacientes atendidos com a Covid-19. Trata-se do projeto de pesquisa "Plataforma Clínica Global sobre a Covid-19 para caracterização clínica e manejo de pacientes hospitalizados com suspeita e confirmação de Covid-19", que tem como pesquisador proponente Fernando Anschau e a instituição proponente o Hospital Nossa Senhora da Conceição, em Porto Alegre (CAAE 41610920.1.1001.5530).

É um desafio bastante complexo se consideradas as heterogeneidades regionais, assistenciais e de modelos que regem tal assistência, de registro e armazenamento de informações clínicas e epidemiológicas em prontuários eletrônicos de instituições hospitalares com perfis tão distintos. Além disso, os dados analisados são protegidos pelos aspectos éticos e de proteção dos direitos dos participantes de pesquisa, visto tratar-se de informações sensíveis, condição que tornou necessária a definição e implantação de modelo de governança de dados em que todos os hospitais participantes pudessem: 1. conhecer e acompanhar as diferentes etapas da pesquisa; 2. perceber os pontos de segurança da informação; 3. ter o completo e inequívoco processo de envio dos dados até um repositório central (DataHub ou centro de dados); e 4. processar as informações de maneira ética, segura e em conformidade com a legislação vigente.

De acordo com o The Governance Institute (DGI)[1], governança de dados é um sistema de direitos de decisão e responsabilidades para processos relacionados à informação, executado de acordo com modelos acordados que descrevem quem pode realizar quais ações com quais informações, e quando, em que circunstâncias, usando quais métodos.

A Figura 1 caracteriza a estrutura do comitê de governança de dados criado para responder às demandas de armazenamento, extração, processamento de dados e da curadoria científica do projeto. O nível estratégico é representado pelo Comitê de Assessoramento do Projeto Plataforma OPAS Covid-19 já constituído e que possui como finalidade principal apoiar tecnicamente a Organização Pan-Americana da Saúde nos termos da Plataforma Global de Dados Clínicos Covid-19 e instituições participantes para a elaboração de análises e publicações dos dados clínicos e epidemiológicos. No nível tático, encontram-se os comitês e atores responsáveis pela tradução da estratégia delimitada pelo nível estratégico, nível este em que são propostas as soluções que viabilizam os objetivos traçados e onde se encontra a gestão do projeto. O nível operacional executa as atividades oriundas do nível tático sempre em observância aos padrões e regras de segurança de dados.



1. The Governance Institute. Disponível em: https://bit.ly/3URnUOh. Acessado em 20/11/2022.

Figura 1. *Representação esquemática do Comitê de Governança de Dados do projeto*

O Comitê de Governança de Dados criou um DataHub (centro de armazenamento de dados) para ser o repositório eletrônico de dados assistenciais dos hospitais coparticipantes do projeto de pesquisa. O DataHub está armazenado na nuvem, em empresa de processamento de dados, em conformidade com os padrões estabelecidos pela Lei de Portabilidade e Responsabilidade de Seguro Saúde (Safe Harbor e Expert Determination) e pelo Regulamento Europeu Geral de Proteção de Dados (GDPR 2016/679).

Prevê processos de desidentificação e anonimização necessários para assegurar privacidade e confidencialidade das informações de participantes de pesquisa, nos termos estabelecidos pela da legislação norte-americana para proteção de dados em saúde – Health Insurance Portability and Accountability Act (HIPAA)[2] – e nacional quanto à Lei Geral de Proteção de Dados Pessoais (LGPD)[3]. Os processos de desidentificação contemplam a remoção de dados considerados protegidos, sendo eles: nomes, endereços (incluindo código postal), todas as datas e informações de contato (e-mail, telefone), além de fotos que possam identificar as pessoas. Tais processos visam a assegurar privacidade, permitindo a utilização das informações em saúde de maneira segura com relação aos aspectos de privacidade e sigilo.

A preparação do dataset do projeto previu as etapas de: a) acesso e convergência de bancos de dados assistenciais dos hospitais participantes de diferentes fontes para a construção de uma única base de dados assistencial; b) identificação da qualidade dos dados, incluindo a formatação, disponibilidade, presença de valores faltantes ou discrepantes/duplicatas e a criação de novos campos por meio de algoritmos de mineração de dados em campos não

[2]*Health Insurance Portability and Accountability Act (HIPAA). Disponível em: https://bit.ly/3hXzKbc. Acessado em 21/11/2022.*

[3]*Disponível em: https://bit.ly/3TUFdg6. Acessado em 21/11/2022.*

não estruturados; c) combinação de colunas individuais em dados mediante a elaboração de medidas de tendências centrais e de dispersão, valores máximos e mínimos, permitindo a caracterização dos dados; d) estruturação dos dados, tornando-os passíveis de análise estatística, e a caracterização de achados, padrões e formulação de hipóteses que permitam análise visual ou modelagem estatística; e) limpeza de dados: nomeação/renomeação de colunas de variáveis; f) identificação e substituição de variáveis com problemas no registro; g) agregação e combinação de variáveis, incluindo remoção de duplicatas; h) manipulação dos dados: programação e utilização de filtros, combinação de bases de dados; i) reformatação dos datasets: transposição, desdobramentos, inserções e substituições; j) beneficiamento dos dados: processamento de dados não estruturados (ver adiante); l) elaboração do produto final dos dados: estatística descritiva (medidas de tendência central e dispersão) e gráficos. Para a preparação do dataset do projeto, foi necessária a utilização de plataforma web para permitir a sequência de etapas acima descrita, a visualização delas e a possibilidade de uso de estratégias de aprendizagem de máquina.

Os pesquisadores utilizaram várias atividades de processamento dos dados por meio do desenvolvimento de códigos de programação mediante o uso de softwares R e Rstudio a fim de combinar os diferentes módulos correspondentes às variáveis estruturadas extraídas diretamente das bases de dados hospitalares. Para reunir as internações equivalentes, por exemplo, foi utilizada a variável correspondente ao código de atendimento, sendo que esse processo ocorreu em etapas sequenciais, com conferência dos totais de casos em cada passo com bases externas. Foram feitas distintas estratégias e etapas de avaliação e análise de correspondências comparativas entre os processos de extração automatizada dos bancos de dados e análises realizadas por médicos e enfermeiros revisando manualmente os registros clínicos dos prontuários eletrônicos, usados como parâmetro de referência.

Para que as etapas de processamento acima sumarizadas fosse possível, foi identificada a necessidade de uso de plataforma eletrônica capaz de reunir essa sequência de operações. Dessa forma, a partir da intercessão de analista sênior de tecnologia da

informação do HCPA, este consultor solicitou e obteve licença "acadêmica" (sem ônus de uso) para utilização dos serviços da plataforma Dataiku para ações de curadoria científica dos dados do projeto.

A empresa Dataiku[4] é atualmente a plataforma líder para soluções de inteligência artificial empregadas para o cotidiano de empresas, organizações e sistemas. Reúne diversas soluções para implantação, uso, gerenciamento e desenvolvimento de análises de dados baseados em ferramentas de inteligência artificial utilizando componentes pré-construídos e processos automatizados, sempre que possível, para evidenciar linhas de fluxos de processos de trabalho, assim como estratégias efetivas de gerenciamento e governança entre equipes de maneira a criar programas de análise de dados transparentes, reprodutíveis e escaláveis utilizando inteligência artificial.

A partir da preparação do dataset do projeto, passou-se à elaboração e avaliação do modelo de análise dos dados, para posterior elaboração do DataHub, extração dos relatórios de dados, interpretação do modelo e geração das análises estatísticas. Dessa forma, para as etapas de elaboração de avaliação do modelo de análise, foi possível verificar a factibilidade do software/plataforma utilizado no projeto, como será caracterizado nas etapas a seguir.

Considerando que as informações contidas nos registros hospitalares de cada instituição podem ser divididas entre estruturadas e texto aberto (ou seja, não estruturado), e tendo em vista as heterogeneidades dos hospitais brasileiros em termos das práticas assistenciais e de gestão, além do próprio registro e armazenamento de informações clínicas, foi necessária a organização da coleta de dados de modo a integrar as informações em texto aberto por meio de interface de programação automatizada.

A programação automatizada aplicada no projeto foi desenvolvida por profissionais de tecnologia da informação para coletar e compartilhar dados de bancos de dados de diferentes hospitais.

---

[4]. *www.dataiku.com*

Esses bancos de dados, com disposição para armazenamento seguro baseado em nuvem, contêm registros de saúde eletrônicos sem a identificação dos participantes de pesquisa no ambiente de rede da própria instituição, correspondentes às notas de admissão nas primeiras 48 horas de atendimento e até 24 horas antes da alta/desfecho de cada participante de pesquisa.

A coleta de dados clínicos sem identificação permitiu reunir: a) as principais características clínicas e fatores prognósticos dos casos de hospitalização por suspeita ou confirmação de Covid-19, ampliando o conhecimento sobre a severidade, espectro e impacto da doença na população hospitalizada globalmente, em diferentes países; b) identificação das intervenções clínicas aplicadas nos atendimentos, dando subsídios para o planejamento operacional global e dos países durante a pandemia de Covid-19.

Os dados estruturados dos prontuários eletrônicos contêm informações demográficas, prescrições, resultados laboratoriais, sinais vitais e caracterização geral dos pacientes, como altura e peso, datas de admissão e desfecho. Como a falta de dados para diferentes variáveis foi importante na primeira fase desse projeto, tornou-se necessário o uso de informações contidas em campos não estruturados, por meio de estratégias de extração em texto aberto, nas evoluções clínicas. Para esse procedimento, foi utilizado o software Smart Health Connect (SHC), que opera com base em algoritmo utilizando redes neurais profundas/"Deep Learning", extraindo informações das evoluções clínicas e incorporando dados aos formulários eletrônicos de pesquisa propostos pela OMS.

O SHC é uma plataforma baseada em nuvem híbrida com instalações federadas em sites. A finalidade do uso do software SHC é melhorar a eficiência na análise de dados não estruturados (dispostos em campos de livre preenchimento nos prontuários eletrônicos) e aumentar a eficácia na identificação de informações demográficas, clínicas e laboratoriais de pacientes a partir de registros assistenciais disponíveis em bases de dados assistenciais.

A empresa iHealth[5] , com base em Brasília (DF), concedeu, em agosto de 2020, licença de uso do software do SHC ao Hospital de Clínicas de Porto Alegre (HCPA) com a finalidade precípua de utilização dos dados no âmbito do projeto, mediante contrato de cessão de direitos de uso sem ônus financeiro, na contratação de licença de uso durante a execução do projeto, sem a transferência de direitos de propriedade ou títulos referentes à Propriedade Intelectual para o HCPA, garantindo o cumprimento das exigências de uso do software definidas em contrato, que incluem os procedimentos de privacidade e segurança de dados dos sujeitos de pesquisa. Para tal, os dados foram armazenados em DataHub na nuvem de processamento, permitindo que o software SHC receba, minere e processe os dados de evoluções clínicas dos prontuários de pacientes, devidamente desidentificados e anonimizados, viabilizando a elaboração do DataHub e a extração de dados necessários às análises da pesquisa.

A ferramenta SHC possui algoritmo representado por conjunto de regras que executam em sequência e ordem procedimentos para, no contexto do projeto em questão, identificar dados de interesse da pesquisa que estejam disponíveis em campos de dados estruturados (variáveis a serem preenchidas) e dados não estruturados (campos de texto livre). A sequência de procedimentos prevista pelo algoritmo preparou os dados para a execução do modelo de aprendizagem de máquina, representado por fórmulas ou códigos de computação que assumem os dados como objeto de análise, aplicam as regras do algoritmo utilizado, processam os dados e produzem relatórios como resultado da análise.

O SHC realizou modelos supervisionados e não supervisionados de aprendizagem de máquina, sendo que o primeiro compreende a utilização de um conjunto de variáveis para a predição do valor de variável a ser obtida. Ou seja, valores prévios são usados para predizer novos valores. No segundo, variáveis em um dataset não rotulado são utilizadas para inferências de padrões para compreensão e agrupamento dos dados. Para fins da realização de modelos supervisionados e não supervisionados de aprendizagem de

[5]. *http://www.ihealthgroup.com.br/*

máquina, a ferramenta SHC utilizou redes neurais artificiais que são modeladas à semelhança de redes de neurônios, que interagem visando interpretar informações e resolver problemas, também com deniminação de deep learning.

As principais variáveis de interesse incorporadas por meio desse software são comorbidades, eventos clínicos, tais quais suporte de oxigênio e tipo de ventilação, tipicamente relacionados à Covid-19, e desfechos. O SHC utiliza as mais novas técnicas de processamento de linguagem natural disponíveis à comunidade científica e modelo de arquitetura de software que viabiliza a extração de dados em massa.

A elaboração de modelo de aprendizagem de máquina previu os seguintes estágios: 1) classificação dos objetos para os quais se pretende desenvolver modelos de predição. Por exemplo: como reconhecer sinais e sintomas relacionados às diferentes apresentações clínicas da Covid-19? E esta classificação pode ser binária (duas classes) ou multiclasse; e 2) modelos de regressão capazes de predizer padrões que permitam sua análise e utilização futura.

As metodologias de desenvolvimento de modelos de aprendizagem de máquina sustentam ser necessário o emprego de tempo adequado na análise, exploração e limpeza dos dados não apenas para a obtenção de melhores resultados, mas também para evitar erros, como dados ou análises enviesadas. Portanto, foi necessário interagir com equipes que estão familiarizadas no registro e análise dos dados em questão para garantir que os dados resultantes da análise automatizada correspondem às experiências empíricas das equipes em saúde.