SRAG no Brasil

Desafios de Qualidade de Dados e Predição de Casos Graves - Um Estudo com Dados Nacionais de 2019 a 2025 entre adultos

Andrey Gabriel Ferreira Gonçalves[1], Jaqueline Nobre da Silva[1] e Julia Peghini Vilela Borges[1]

Instituto Federal de Educação, Ciência e Tecnologia de Brasília, Via L2 Norte, SGAN 610 (610 Norte), Módulo D, E, F e G .Brasília/DF. CEP: 70830-450, Brasil
tsi.cbra@ifb.edu.br
https://www.ifb.edu.br/brasilia/

Resumo A Síndrome Respiratória Aguda Grave (SRAG) é um importante agravo de saúde pública no Brasil, especialmente após a pandemia de COVID-19. Este estudo analisa dados de vigilância da SRAG entre 2019 e 2025, com foco na identificação de fatores associados à admissão em Unidade de Terapia Intensiva (UTI), utilizada como *proxy* de gravidade clínica. Dados públicos do OpenDataSUS foram submetidos a etapas de limpeza, padronização e consolidação, restringindo-se a análise a pacientes hospitalizados adultos (18–120 anos). Foram realizadas análises descritivas, temporais e correlacionais, além da aplicação de modelos supervisionados de Regressão Logística e Árvore de Decisão, avaliados por métricas como precisão, *recall*, *F1-score* e *AUC-ROC*. Os resultados indicam que idade e comorbidades, especialmente obesidade e doença renal crônica, predominam a admissão em UTI. Os modelos apresentaram desempenho moderado, compatível com bases ruidosas e desbalanceadas, sendo mais adequados para análises exploratórias do que para decisão clínica automatizada. O estudo evidencia o potencial analítico da base SRAG, bem como a influência das limitações de qualidade dos dados na modelagem preditiva.

Palavras-chave: SRAG · COVID-19 · Árvore de Decisão · Regressão logística

1 Introdução

A Síndrome Respiratória Aguda Grave (SRAG) é uma condição caracterizada por insuficiência respiratória aguda decorrente de agentes infecciosos, intoxicações ou agravos respiratórios de diversas origens [1]. Desde 2009, o Ministério da Saúde mantém um sistema estruturado de vigilância de SRAG, inicialmente motivado pela pandemia de influenza A (H1N1) e posteriormente expandido para abarcar outros vírus respiratórios relevantes [2].

Em dezembro de 2019, foi identificado na cidade de Wuhan, China, um novo coronavírus, causada pelo agente etiológico severe acute respiratory syndrome

coronavirus-2 (SARS-CoV-2), agente etiológico da Coronavirus 2019 (COVID-19) [3]. A COVID-19 é uma doença infecciosa caracterizada principalmente por febre, tosse, fadiga e sintomas respiratórios, podendo evoluir para formas graves, incluindo pneumonia viral, hipóxia e falência respiratória [4]. A rápida disseminação global levou a Organização Mundial da Saúde a declarar pandemia em março de 2020.

No Brasil, a COVID-19 impactou profundamente o sistema de vigilância e de atenção à saúde. A base SRAG tornou-se o principal instrumento nacional para monitoramento de internações, gravidade, perfil demográfico e evolução clínica dos casos de COVID-19 hospitalizados. Entre 2020 e 2022, observou-se um aumento abrupto no volume de registros, acompanhando as diferentes ondas pandêmicas, influenciadas pela circulação de variantes (Gama, Delta, Ômicron) e pelas desigualdades regionais de acesso ao sistema de saúde [5] [6]

A partir de 2021, o avanço da vacinação modificou o panorama epidemiológico, reduzindo internações e mortalidade nas populações vacinadas, e alterando o perfil de gravidade dos casos registrados [7]). Ainda assim, parte expressiva dos casos de SRAG hospitalizados continuou evoluindo para quadros críticos, exigindo UTI, ventilação mecânica e suporte avançado.

Apesar de sua relevância, a base SRAG apresenta desafios estruturais: heterogeneidade no preenchimento, alta frequência de valores ignorados, inconsistências temporais e mudanças nos formulários ao longo dos anos [8] [9]. Esses fatores reforçam a necessidade de processos rigorosos de limpeza, integração, harmonização e análise dos dados.

Neste contexto, o presente estudo tem como objetivos:

- Consolidar dados SRAG do período de 2019 a 2025;
- Avaliar a qualidade e consistência das variáveis;
- Realizar análises descritivas e analíticas, incluindo avaliação de comorbidades, distribuição temporal e aspectos clínicos;
- Idade adulta válida 18 a 120;
- Identificar fatores associados à admissão em Unidade de Terapia Intensiva (UTI);
- Treinar um modelo preditivo baseado em Regressão logistica e Árvore de Decisão.

A admissão em UTI foi utilizada como proxy de gravidade clínica, ou seja, como uma medida substituta que representa casos com evolução mais severa. Em bases de vigilância como o SRAG, variáveis clínico-laboratoriais (por exemplo, saturação de oxigênio, necessidade de ventilação mecânica ou parâmetros respiratórios) apresentam elevado número de campos vazios, preenchimento inconsistente ou ausência nos anos iniciais da pandemia. Assim, indicadores diretos de gravidade nem sempre podem ser utilizados em análises longitudinalmente comparáveis.

Por outro lado, o registro de admissão em UTI mostra maior consistência, padronização e completude ao longo dos anos, sendo amplamente empregado em estudos epidemiológicos como indicador robusto de severidade da COVID-19

e de outras infecções respiratórias agudas. Diversas análises nacionais demonstram que a UTI está fortemente associada à progressão clínica desfavorável, complicações críticas e maior risco de óbito [10]. Dessa forma, sua adoção como variável-alvo neste estudo permite identificar casos graves de maneira mais estável e confiável dentro do conjunto SRAG 2019–2025.

2 Materiais e Métodos

2.1 Fonte de Dados

Foram utilizados os datasets públicos disponibilizados no OpenDataSUS, SRAG 2019–2025 (INFLUD19 a INFLUD25) [11]

O volume inicial foi de 4.423.757 notificações.

2.2 Escopo da Análise

Para evitar viés de subnotificação e inconsistências de gravação em casos não hospitalizados, o escopo foi restrito a:

- Pacientes hospitalizados (representando 4.201.819 casos)
- Pacientes hospitalizados que tiveram alta ao final do tratamento (representando 2.658.607 casos, que equivale á 64.7% do total de pacientes hospitalizados)
- Notificações com idade adulta (representando 2.109.536)

2.3 Engenharia e Tratamento dos Dados

O *pipeline* de pré-processamento incluiu as seguintes etapas principais:

- **Padronização de Variáveis:**
  - Concatenação dos arquivos anuais;
  - Normalização de campos categóricos (Comorbidades) conforme a padronização 2019–2025;
  - Unificação dos códigos binários
    - 1 = Sim;
    - 2/9/NA/0 = Não
- **Tratamento de Dados Ausentes (*Missing*):**
  - Identificação de 59–66% de valores ignorados ou não preenchidos nas variáveis de comorbidade;
  - **Adoção da premissa:** ausência de marcação "Sim" resulta em classificação como "Não".
- **Variável de Controle (Idade):**
  - Conversão da unidade para anos;
  - Remoção de idades fora do escopo (menor de 18 ou acima de 120 anos);
  - Análise de distribuição (Análise de tendência central e dispersão, Bloxplot IQR e histograma).

- **Variável Alvo (ALVO_GRAVIDADE – UTI):**
  - **Grave:** Paciente admitido em UTI, que foi descrito como = 1;
  - **Não grave:** Paciente sem admissão registrada, que foi descrito como = 0.
- **Balanceamento (Undersampling Manual):**
  - Redução aleatória da classe majoritária (UTI = 0) para igualar-se à quantidade da classe minoritária (UTI = 1).
  - Base balanceada passou a totalizar 930,196 notificações com as classes possuindo a mesma proporção na base (50-50).

2.4 Análises Estatísticas e Gráficas

Foram conduzidas três frentes principais de análise exploratória:

- **Análise Temporal:**
  - Evolução anual do número de casos;
  - Taxa de admissão em UTI ao longo dos anos;
  - Vizualização das ondas da pandemia (2020, 2021 e 2022).
- **Perfil de Gravidade:**
  - Histograma de idade segmentado por UTI;
  - Boxplots de idade *vs.* gravidade;
  - Taxas de UTI por comorbidade (diabetes, obesidade, doença renal, asma, imunodepressão, síndrome de Down, etc.).
- **Correlações:**
  - Correlação de *P*earson (mede relação linear - reta) e *S*pearman (mede ordem - ranking) entre idade, comorbidades e variável-alvo;

2.5 Modelo de Machine Learning

Foram treinados dois modelos supervisionados para a tarefa de predição de gravidade (admissão em UTI): **Árvore de Decisão** e **Regressão Logística**. A utilização de ambos permitiu comparar como eles lidavam com os desafios da falta de linearidade dos dados e os ruídos no preenchimento das comorbidades.

- **Regressão Logística**: Utilizada como modelo base por ser:
  - Amplamente empregada em problemas clínicos de classificação binária;
  - Adequada para estabelecer relações lineares entre preditores e o desfecho;
- **Árvore de Decisão**: Utilizada como modelo mais adequado por ser:
  - Apropriada para variáveis categóricas e binárias;
  - Robusta a bases ruidosas e desbalanceadas.

Variáveis de entrada (features)

- Idade
- Cardiopatia
- Diabetes
- Obesidade
- Doença Renal
- Asma
- Imunodepressão
- Síndrome de Down
- Doença Hepática Crônica
- Doença Hematológica Crônica
- Outras Pneumopatia Crônica
- Doença Neurológica Crônica

Métricas avaliadas

- F1-Score
- AUC-ROC
- Precisão
- Recall

Os valores numéricos serão apresentados durante a próxima sessão, com maior detalhamento e análise feita durante a pesquisa.

3 Resultados

3.1 Consistência e Qualidade dos Dados

A inspeção inicial dos dados brutos revelou os seguintes desafios:

- **Alta taxa de nulos:** Observada nas comorbidades, superando 60% em na maior parte das variáveis;
- **Preenchimento inconsistente:** Identificado em variáveis clínicas (ex.: a variável IMUNODEPRE apresenta $\approx 67\%$ de dados nulos/ignorados);

Apesar das limitações, a consolidação permitiu formar um *dataset* funcional para a modelagem e as análises subsequentes.

Além disso, optou-se por realizar um recorte etário incluindo apenas pacientes adultos. Essa decisão teve como objetivo reduzir viés associado à heterogeneidade clínica entre crianças e recém nascidos, garantindo maior comparabilidade entre os grupos e maior estabilidade estatística nas análises inferenciais e nos modelos preditivos.

Essas limitações estruturais e de completude dos dados impactam diretamente a capacidade explicativa e preditiva dos modelos, reforçando o caráter exploratório e interpretável das análises realizadas neste estudo.

3.2 Análise Temporal

A curva temporal, vizualizada na Figura 1, confirmou a presença de três grandes ondas da pandemia:

- **2020:** Primeira onda, caracterizada por um impacto moderado;
- **2021:** O pior ano em volume de casos e taxa de admissão em UTI, correlacionado com a circulação de variantes mais agressivas;
- **2022:** Redução gradual de casos graves, coincidente com o avanço da vacinação.

Observou-se que:

- Durante picos de casos, a **taxa de internação em UTI diminuía**, possivelmente devido à sobrecarga hospitalar.
- Em períodos de estabilidade, a taxa voltava a valores mais altos e consistentes.

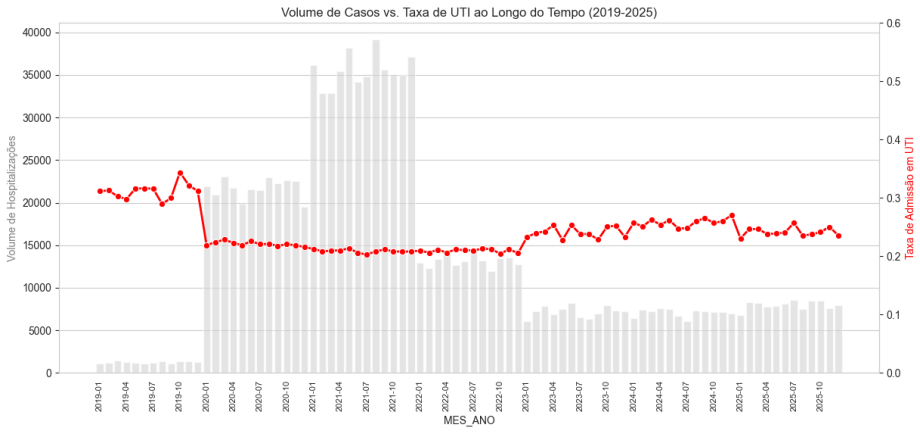

Fig. 1. Volume de hospitalizações vs. Taxa de UTI ao longo dos anos

Embora a análise temporal permita identificar padrões agregados e hipóteses plausíveis sobre sobrecarga do sistema de saúde, ressalta-se que os dados observacionais não permitem inferir causalidade direta entre o aumento de casos e a redução da taxa de admissão em UTI.

3.3 Idade como Fator de Gravidade

O boxplot, da Figura 2 mostra a diferença de distribuição entre os grupos:

- Pacientes em UTI → média de idade maior;

– Faixa etária acima de 60 anos concentra a maior parte dos casos graves como é possivel visualizar na Figura 3 .

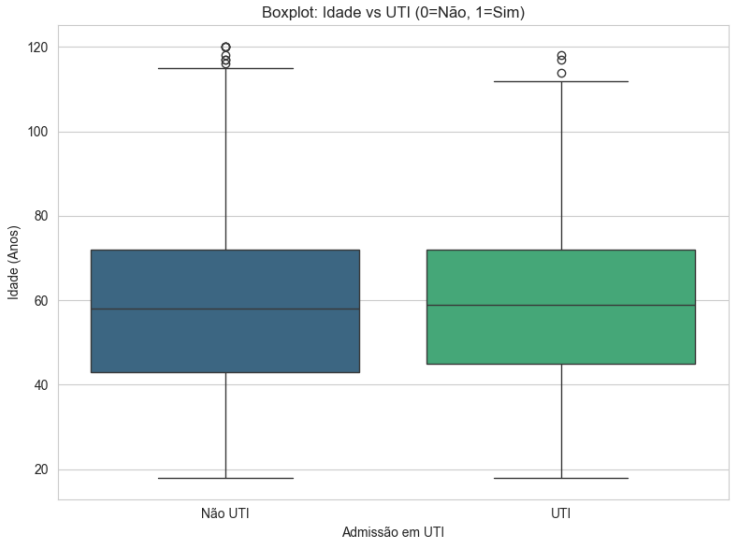

Fig. 2. Idade dos admitidos na UTI vs. fora da UTI

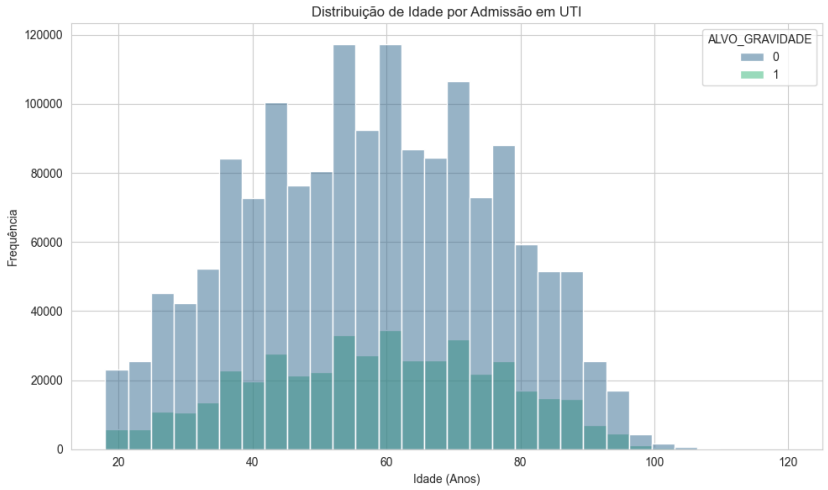

Fig. 3. Distribuição de Idade por Admissão em UTI

A correlação entre idade e gravidade manteve-se positiva, em concordância com a análise preliminar baseada no coeficiente de correlação de Pearson entre idade e admissão em UTI (r = 0,0045). Embora de magnitude reduzida, esse coeficiente indica uma tendência consistente de aumento do risco de internação em UTI com o avanço da idade.

Tal associação permaneceu relevante mesmo após o recorte para a população adulta, sugerindo que o risco de admissão em UTI cresce de forma progressiva ao longo desse intervalo etário. Esses resultados reforçam a idade como o fator isolado mais fortemente associado à gravidade clínica no conjunto de dados analisado

Em termos epidemiológicos, a idade funciona como um marcador agregado de risco, refletindo tanto a maior prevalência de comorbidades quanto a vulnerabilidade fisiológica associada ao envelhecimento.

3.4 Comorbidades e Risco de UTI

A análise percentual revelou (como pode ser observada na Figura 4):

- **Obesidade** → uma das comorbidades com maior proporção de UTI.
- **Doença Renal Crônica** → alta taxa relativa de gravidade.
- **Diabetes e Cardiopatia** → associação moderada, porém consistente.
- **Asma** → associação mais fraca.

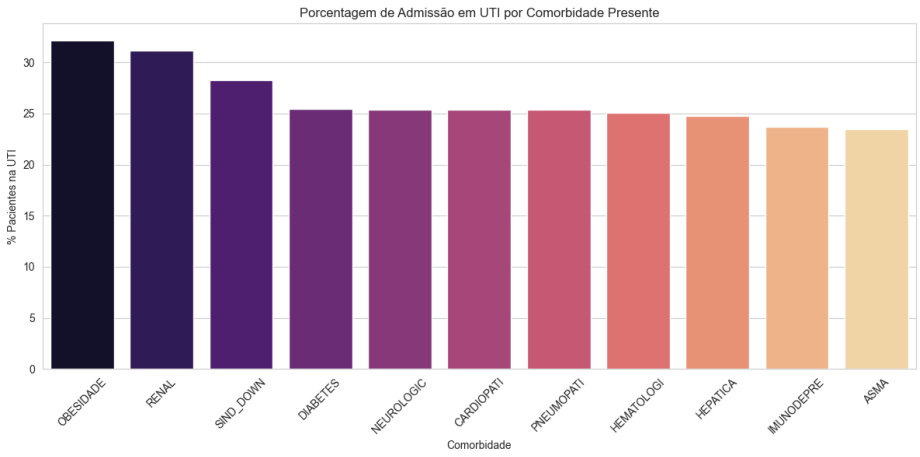

Fig. 4. Porcentagem de admissão na UTI por comorbidade

As associações identificadas entre comorbidades e admissão em UTI devem ser interpretadas de forma observacional, uma vez que o desenho do estudo e as limitações de completude da base não permitem estabelecer relações causais diretas.

3.5 Modelos Preditivos (Árvore de Decisão e Regressão Logística)

Foram avaliados dois modelos supervisionados: Árvore de Decisão e Regressão Logística. Ambos foram avaliados utilizando as métricas de desempenho **precisão**, **recall**, **F1-Score** e **AUC-ROC**, com foco no desfecho de UTI, permitindo uma análise complementar do desempenho geral, da capacidade de detecção de casos graves e da discriminação entre classes.

Os modelos foram treinados e avaliados exclusivamente sobre a população adulta, assegurando maior coerência clínica e reduzindo distorções decorrentes da heterogeneidade entre perfis pediátricos, jovens e idosos extremos(>120 anos). Essa delimitação contribuiu para maior estabilidade dos coeficientes da Regressão Logística e para regras de decisão mais consistentes na Árvore de Decisão.

Dado o caráter desbalanceado da base e a presença de ruídos nos dados clínicos, os modelos devem ser interpretados como ferramentas de apoio à análise exploratória e à identificação de padrões de risco, e não como sistemas automatizados de tomada de decisão clínica.

Regressão Logística

A Regressão Logística foi o primeiro modelo utilizado e foi escolhido como modelo comparativo, dada sua ampla aplicação em cenários de risco clínico. Os resultados anteriores às técnicas de balanceamento e redução de dimensionalidade dos dados indicaram:

- Acurácia geral de 77%, a qual é enganosa por conta da quantidade imensa de falso negativos na classe referente à UTI;
- Precisão (25%);
- *R*ecall (0%); *F1-Score* (0%) muito baixos para a classe de UTI;
- Curva *AUC-ROC* de 55%.

Esses resultados mostraram o grande impacto negativo do viés da base de dados por conta das informações que não foram preenchidas durante o período da pandemia, provavelmente, em casos de subnotificação.

Após o balanceamento dos dados, o modelo de Regressão Logística foi treinado novamente e apresentou os seguintes resultados:

- Acurácia geral de 54%;
- Precisão (55%), *R*ecall (0.50) e *F1-Score* (0.52) aumentaram mas ainda se mantém menores do que o esperado para apoiar à decisão clínica de forma robusta;
- Curva *AUC-ROC* de 0.56.

Esse comportamento está alinhado ao que Grus [12] descreve sobre modelos lineares aplicados a cenários clínicos, nos quais a estabilidade dos coeficientes tende a aumentar à medida que se reduz a heterogeneidade populacional — como no recorte adotado neste estudo. De forma convergente, o trabalho de

Salles[13] também reporta melhor desempenho preditivo em populações adultas, reforçando a coerência dos resultados obtidos.

Ainda que a Regressão Logística apresente maior estabilidade e métricas de desempenho muito boas permanece fortemente condicionado às limitações estruturais da base, indicando que avanços na qualidade e completude dos dados tendem a produzir ganhos mais relevantes do que a escolha do algoritmo em si, por essa razão foi decidido usar mais um modelo, a Arvore de decisão para identificar o desemprenho de um modelo diferente com a mesma base de dados.

Árvore de Decisão

Nesta etapa, foram aplicadas técnicas de balanceamento e redução de dimensionalidade, permitindo observar que o modelo apresentou:

- Redução de Acurácia Geral, após o balanceamento das classes a precisão das predições ficou próxima de 55% para cada uma delas;
- Precisão (55%), *Recall* (0.50) e *F1-score* (0.52) aumentaram mas ainda se mantém menores do que o esperado para apoiar à decisão clínica de forma robusta;
- Curva *AUC-ROC* de 0.56

A análise interpretativa da árvore indicou que, mesmo após o recorte para a população adulta, a idade permaneceu como principal variável de divisão, seguida de comorbidades como obesidade e doença renal crônica, que aparecem de forma recorrente nos ramos superiores do modelo. Esse comportamento é consistente com os princípios de interpretabilidade também descritos por Grus [12], segundo os quais árvores de decisão tendem a privilegiar, nos níveis mais altos da hierarquia, atributos com maior poder discriminativo.

A predominância dessas variáveis também converge com evidências de estudos clínicos sobre COVID-19 e SRAG, que identificam idade avançada, obesidade e doença renal como fatores fortemente associados ao risco aumentado de evolução para admissão em UTI em adultos.

Apesar da interpretabilidade do modelo, seu desempenho limitado evidencia que a qualidade e a completude das variáveis clínicas disponíveis impõem restrições relevantes à capacidade de generalização das regras de decisão aprendidas.

4 Considerações Finais

Este estudo tornou possível visualizar a larga escala da vigilância de SRAG, analisar padrões associados à gravidade clínica e aplicar modelos preditivos interpretáveis sobre a população adulta. A base de dados referente ao período de 2019 a 2025 mostrou-se rica em volume e diversidade, porém desafiadora, exigindo esforços substanciais de padronização, limpeza e harmonização temporal.

Entre os principais achados, destacam-se:

- A idade manteve-se como o fator isolado mais fortemente associado à admissão em UTI, mesmo após o recorte para a população adulta, atuando como marcador agregado de risco clínico;
- O ruído encontrado nas variáveis específicas de comorbidade, como obesidade e doença renal crônica, tornaram inviável uma predição consistente de admissão em UTI baseado nesse fator;
- As queda de casos de admissão em UTI durante as ondas da pandemia sugere uma grande possibilidades de viés e ruído gerados por conta de subnotificação;
- Os modelos utilizados apresentaram desempenho compatível com bases ruidosas; contudo, não foi possível avaliar de forma conclusiva o *proxy* de gravidade adotado. Ainda assim, os resultados mostraram-se úteis para análises exploratórias, embora não necessariamente conclusivas.

Os resultados reforçam a relevância da vigilância epidemiológica baseada em dados abertos e evidenciam que melhorias na qualidade, completude e padronização das variáveis clínicas tendem a produzir ganhos mais expressivos do que a simples adoção de algoritmos mais complexos. Nesse sentido, os modelos avaliados devem ser compreendidos como ferramentas de apoio à análise e ao planejamento em saúde, e não como sistemas automatizados de decisão clínica.

Como trabalhos futuros, sugere-se a incorporação de variáveis adicionais, como informações laboratoriais, desfecho clínico final, bem como a avaliação de modelos mais robustos para dados desbalanceados. Tais avanços podem contribuir para o desenvolvimento de ferramentas mais precisas de apoio à vigilância e à tomada de decisão em contextos de crises sanitárias.

Referências

1. Brasil, Ministério da Saúde: Vigilância Sentinela de Síndrome Gripal e Vigilância de Síndrome Respiratória Aguda Grave. Secretaria de Vigilância em Saúde, Brasília (2014)
2. Brasil, Ministério da Saúde: Protocolo de Vigilância da Síndrome Respiratória Aguda Grave (SRAG). Ministério da Saúde, Brasília (2009)
3. Jin, Y., Yang, H., Ji, W., Wu, W., Chen, S., Zhang, W., Duan, G.: Virology, epidemiology, pathogenesis, and control of COVID-19. Viruses 12(4), 372 (2020). https://doi.org/10.3390/v12040372
4. World Health Organization (WHO): Coronavirus disease (COVID-19): Situation Report – 51. WHO, Geneva (2020)
5. Bastos, L.S., et al.: COVID-19 e a sobrecarga dos sistemas de vigilância no Brasil: desafios e perspectivas. Cadernos de Saúde Pública 37(3) (2021)
6. Brasil, Ministério da Saúde: Boletim Epidemiológico – COVID-19. Ministério da Saúde, Brasília (2021)
7. Hitchings, M.D.T., et al.: Effectiveness of CoronaVac among healthcare workers in a setting of high SARS-CoV-2 Gamma variant transmission in Brazil. The Lancet Regional Health – Americas 1 (2021)
8. De Paula, R.C., et al.: Qualidade dos dados de Síndrome Respiratória Aguda Grave no Brasil durante a pandemia. Revista Epidemiologia e Serviços de Saúde 31(4) (2022)

9. DATASUS, Ministério da Saúde: Base de Dados de Síndrome Respiratória Aguda Grave (SRAG). DATASUS, Brasília (2023)
10. Ribeiro, K.B., et al.: Clinical severity of COVID-19 in Brazil: factors associated with ICU admission and mortality. BMC Infectious Diseases 21 (2021)
11. DATASUS, Ministério da Saúde: Banco de Dados de Síndrome Respiratória Aguda Grave (SRAG) 2021–2024. Disponível em: https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024. Acesso em: 09 dez. 2025.
12. Grus, J.: *Data Science do Zero*. Alta Books, Rio de Janeiro (2021). E-book. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788550816463/. Acesso em: 20 nov. 2025.
13. Salles, D. B.: Predição de desfecho desfavorável em pacientes com COVID-19 usando técnicas de aprendizado de máquina. Monografia (Graduação em Ciência da Computação), Universidade Federal de Ouro Preto, Instituto de Ciências Exatas e Biológicas (2023). Disponível em: https://www.monografias.ufop.br/bitstream/35400000/5902/8/MONOGRAFIA$_{P}redioPacientesCOVID$19.pdf.$Acessoem$: 20nov.2025.