TECH CHALLENGE – FASE 03

ANÁLISE DOS INDICADORES POPULACIONAIS E CLÍNICOS DA COVID-19

Bruna Alves de Amorim

Matheus Cesar do Amaral

Donizeti Carlos dos Santos Junior

Yuri Tierno Popic

## **1 Introdução**

A pandemia de COVID-19 constituiu um dos maiores desafios sanitários da história recente, afetando não apenas os sistemas de saúde, mas também as estruturas sociais e econômicas em escala global. No Brasil, a progressão dos casos foi fortemente influenciada por desigualdades estruturais que condicionaram o acesso desigual à assistência médica e a serviços de diagnóstico. Segundo o Núcleo de Operações e Inteligência em Saúde, a taxa de letalidade no país mostrou-se elevada, influenciada diretamente pelos determinantes sociais e pela desigualdade no acesso ao tratamento hospitalar.

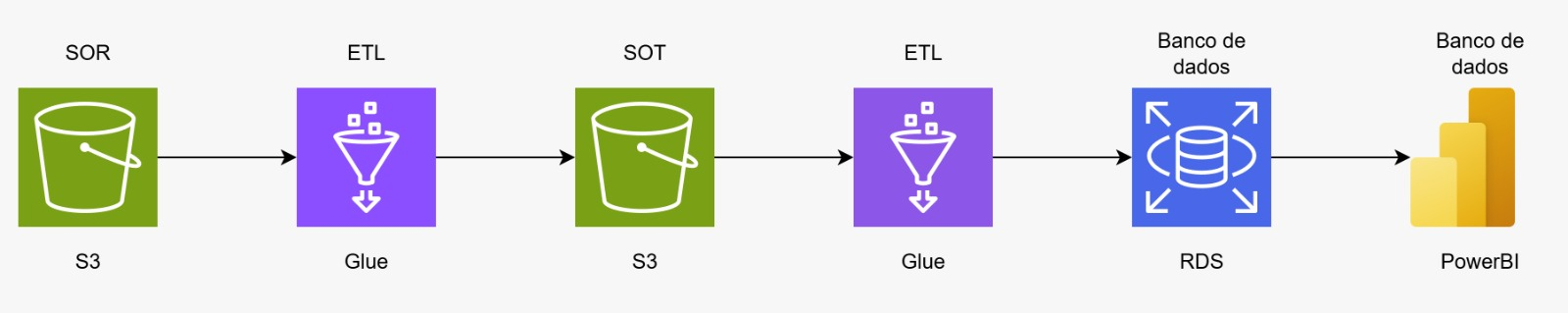
Além dos fatores epidemiológicos, o contexto pandêmico forçou uma mudança abrupta no modelo tradicional de atendimento médico. As instituições de saúde tiveram de abandonar, temporariamente, o padrão de cuidado presencial e investir em soluções tecnológicas que permitissem o acompanhamento clínico remoto dos pacientes. Essa transição representou um duplo desafio para os profissionais de saúde: compreender uma nova doença e, simultaneamente, adaptar-se a novas formas de prestação de cuidado.

Nesse cenário, o avanço das tecnologias digitais aplicadas à saúde — como teleconsultas, plataformas de triagem automatizada e monitoramento remoto — emergiu como alternativa efetiva e segura para manter a comunicação entre pacientes e profissionais. A pandemia, portanto, consolidou-se como um marco de transformação tecnológica no setor, impondo a necessidade de novas estratégias assistenciais e de adequação dos serviços à realidade do distanciamento social.

O presente estudo, desenvolvido sob a perspectiva da análise de dados, tem como objetivo compreender o comportamento populacional durante a pandemia da COVID-19 no Brasil, com base nos microdados da **PNAD-COVID-19** do Instituto Brasileiro de Geografia e Estatística (IBGE). Pretende-se identificar padrões clínicos, econômicos e sociais relevantes para subsidiar o planejamento hospitalar e orientar ações estratégicas em caso de novos surtos epidêmicos.

## **2 Metodologia**

O processo foi estruturado em um **pipeline de ingestão e transformação sobre a AWS**, conforme ilustrado na **Figura 1**, que apresenta o fluxo completo desde a ingestão dos dados brutos até a camada analítica no Power BI.



Como é possível ver na imagem, o fluxo inicia-se na **camada SOR (System of Record)**, hospedada no Amazon S3, onde os arquivos CSV originais são armazenados de forma íntegra. Em seguida, o **AWS Glue** executa o primeiro processo de **ETL (Extract, Transform, Load)**, convertendo os arquivos brutos em formato **Parquet** e gravando-os na camada **SOT (System of Truth)**.

Esse job é parametrizado com variáveis como **SOR\_PATH**, **SOT\_PATH**, **DELIMITER**, **HAS\_HEADER**, **INDEX\_MAP** e **FAIL\_ON\_MISSING**, garantindo flexibilidade e robustez para lidar com diferentes formatos de entrada.  
 Na leitura, o job identifica automaticamente a presença ou ausência de cabeçalhos, ajusta o delimitador e aplica um comportamento *fail-fast* caso detecte inconsistências estruturais, interrompendo a execução quando os metadados esperados não são encontrados. Durante a extração, também são obtidos os valores de ano e mês diretamente do caminho do arquivo, validando a estrutura de partição esperada (“.../ano=YYYY/mes=MM/...”), o que assegura rastreabilidade temporal.

A etapa de transformação contempla uma série de mapeamentos e enriquecimentos semânticos: códigos numéricos de UF são convertidos em siglas estaduais; variáveis categóricas (como **V1022**, **A003** e **A005**) são rotuladas em texto legível; e as respostas codificadas em **1/2/3/9** são padronizadas em “Sim”, “Não”, “Não sabe” e “Ignorado”.  
 A variável de idade é tratada e classificada em faixas etárias (“0–14”, “15–24”, “25–39”, “40–59” e “60+”), compondo um conjunto final de 20 variáveis padronizadas e rotuladas em texto\*, já adequadas ao consumo analítico.

A gravação no **S3/SOT** é feita em **Parquet particionado por ano, mês e uf\_sigla**, com sobrescrita dinâmica e repartition alinhado às mesmas chaves, o que otimiza a leitura e reduz o custo de *scan*. Esse processo conclui a primeira etapa do pipeline (SOR → SOT) e prepara os dados tratados para ingestão relacional.

Na sequência, o job *glue\_sot\_to\_rds.py* executa a segunda etapa (**SOT → RDS**), responsável por carregar todos os registros processados do **S3** (Parquet) para o **Amazon RDS (PostgreSQL)**. O script implementa um carregamento completo (*full load*), sem deduplicação, e valida previamente a integridade dos dados através de contagens por ano e mês antes da gravação, assegurando correspondência com o volume esperado no *data lake*.  
 O comportamento de escrita pode ser configurado via parâmetro **OVERWRITE**: quando ativado, o processo executa um *truncate overwrite* completo; caso contrário, realiza um *append incremental*.

A escrita **JDBC** utiliza parâmetros de desempenho como **BATCHSIZE** e **COALESCE**, otimizando a quantidade de partições e conexões simultâneas com o banco. Além disso, o job força a tipagem final coerente com o modelo do **RDS** (inteiros para ano/mês e texto para as demais colunas), garantindo consistência de *schema*. Essa etapa conclui o fluxo de integração com a base relacional, que servirá de origem para as consultas do **Power BI**, consolidando um ciclo de ingestão totalmente automatizado, rastreável e auditável.

### **2.1 Fonte de dados e ambiente de análise**

As fontes de dados consistem em arquivos **CSV oriundos do PNAD**, armazenados na camada **SOR** do Amazon **S3**, onde permanecem íntegros e disponíveis para reprocessamentos.  
 A arquitetura utiliza exclusivamente serviços gerenciados da **AWS**, compondo um ecossistema escalável e seguro: o **S3** como *data lake* (camadas **SOR** e **SOT**), o **Glue** como motor de **ETL** em **PySpark**, e o **RDS (PostgreSQL)** como camada relacional.

O **Glue Data Catalog** e os *crawlers* mantêm o controle de metadados e esquemas, viabilizando consultas via **Athena** e automação de pipelines.  
 O uso do formato **Parquet** e da partição por chaves temporais e regionais melhora a eficiência das leituras e facilita o *tracking* de versões históricas.

Na camada analítica, o **Power BI** é conectado diretamente ao banco **RDS**, consumindo as tabelas tratadas e modeladas. Dentro do ambiente do Power BI, foram criadas **colunas binárias derivadas das variáveis textuais** exportadas pelo Glue (por exemplo, transformando respostas “Sim”/“Não” em valores 0 e 1).  
 Essas colunas servem de base para a elaboração de medidas **DAX** que calculam proporções, contagens condicionais e indicadores de atendimento ou sintomas, permitindo maior flexibilidade nas análises e visualizações. Esse recurso facilita o cruzamento entre dimensões e fatos, otimizando a construção de **KPIs** e dashboards interativos que refletem a realidade observada nos dados processados.

O dashboard pode ser visualizado através do seguinte link: https://app.powerbi.com/links/IgZ\_h5F9Cj?ctid=11dbbfe2-89b8-4549-be10-cec364e59551&pbi\_source=linkShare.

### **2.2 Variáveis selecionadas**

Atendendo à limitação de vinte variáveis, foram selecionados os seguintes questionamentos da PNAD-COVID-19:

|  |  |
| --- | --- |
| **Dimensão** | **Variáveis selecionadas** |
| **Clínica** | febre, tosse, dificuldade\_respirar, perda\_olfato\_paladar, fadiga |
| **Atendimento e gravidade** | procurou\_atendimento, atendimento\_sus\_hospital, atendimento\_priv\_hospital, internacao, sedado\_entubado |
| **Demográfica e social** | sexo, faixa\_idade, escolaridade, situacao\_domicilio (urbano/rural), uf\_sigla |
| **Econômica** | auxilio\_emergencial, bolsa\_familia, seguro\_desemprego, ocupacao, rendimento |
| **Temporais** | ano, mes |

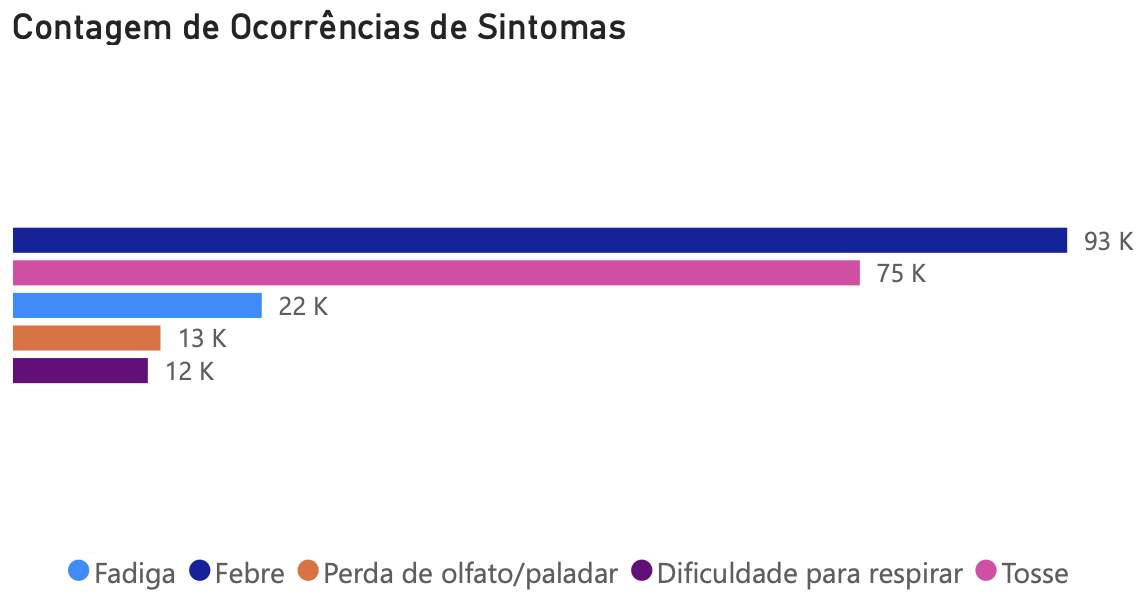
Essas variáveis permitiram a construção de indicadores que refletem tanto a dimensão clínica da doença quanto suas implicações comportamentais e socioeconômicas.

## **3 Resultados**

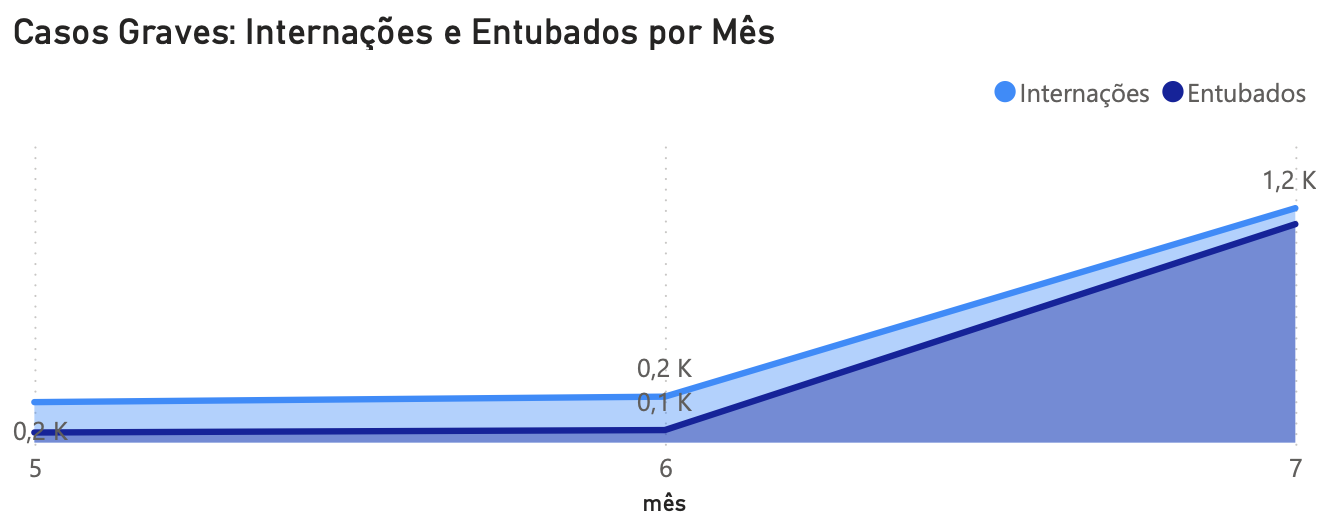
### **3.1 Caracterização clínica da população**



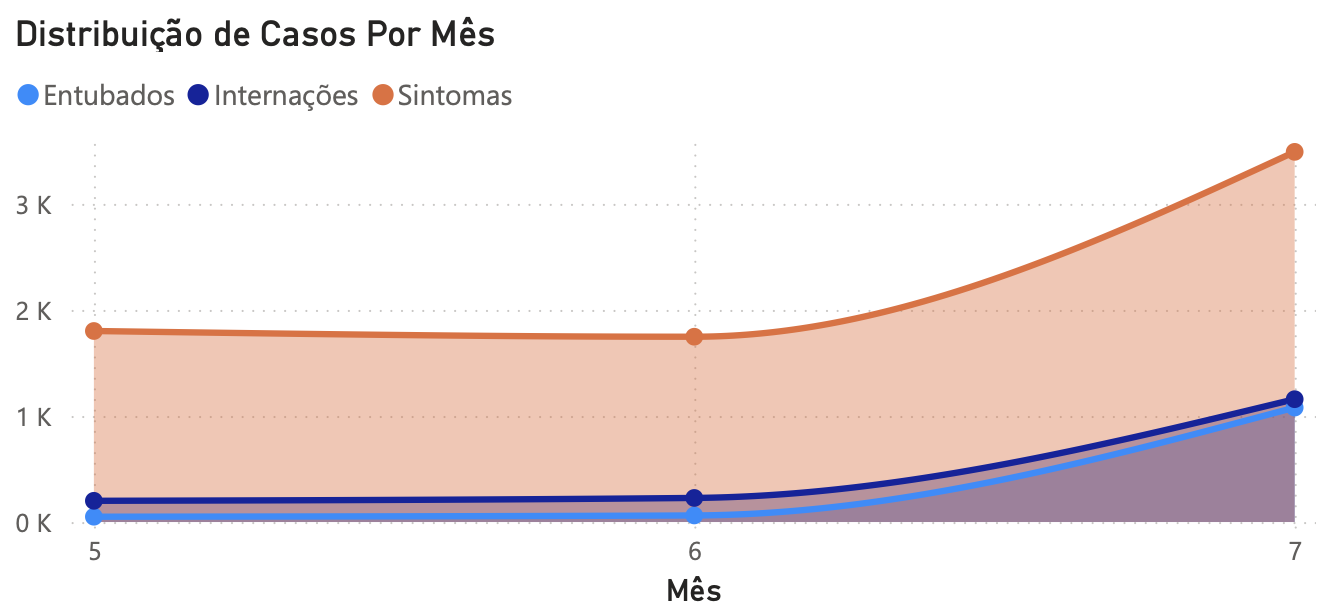
Entre os meses de maio e julho de 2020, aproximadamente **126 mil pessoas** declararam sintomas compatíveis com COVID-19. Destas, **2.776** evoluíram para casos graves, o que representa **2,2%** do total de sintomáticos.



Os sintomas mais prevalentes foram **fadiga (74%)** e **dificuldade para respirar (60%)**, seguidos de **febre (17,5%)**, **tosse (10,3%)** e **perda de olfato/paladar (9,5%)**.  
Esses dados apontam para um quadro clínico predominantemente respiratório, exigindo maior disponibilidade de **oxigenoterapia, ventilação mecânica e suporte de imagem torácica** nas unidades hospitalares.

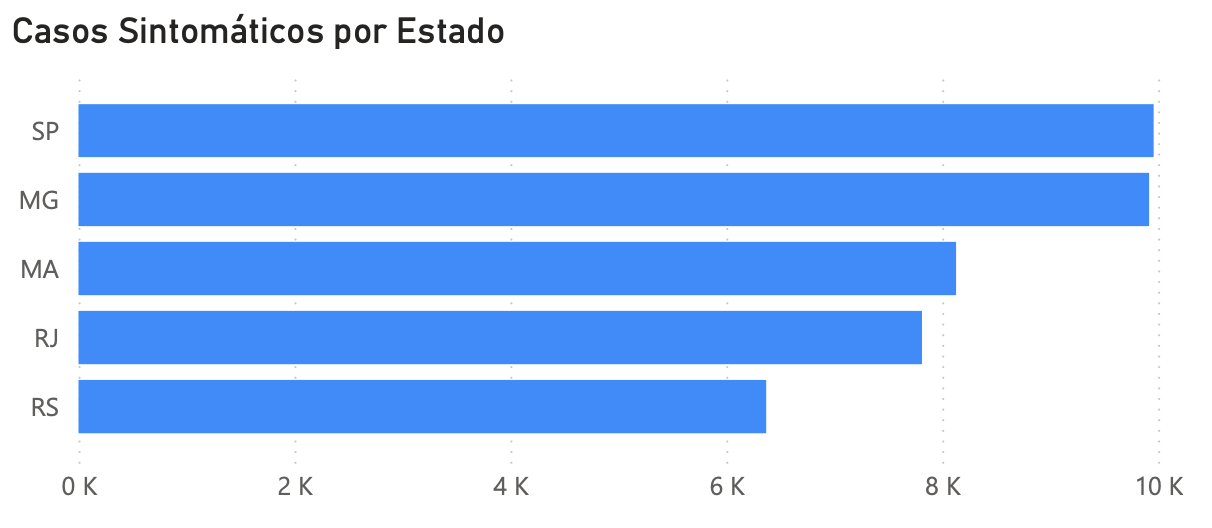


Observa-se um crescimento expressivo tanto nas **internações** quanto nos **casos que evoluíram para entubação** entre os meses de maio e julho. A partir de junho, o número de internações praticamente **quadruplica**, enquanto o de pacientes entubados acompanha a mesma tendência de alta, indicando uma **maior gravidade dos quadros clínicos**. Esse aumento pode estar relacionado à **circulação mais intensa do vírus** e ao **esgotamento da capacidade ambulatorial**, levando mais pacientes a necessitarem de suporte hospitalar avançado.



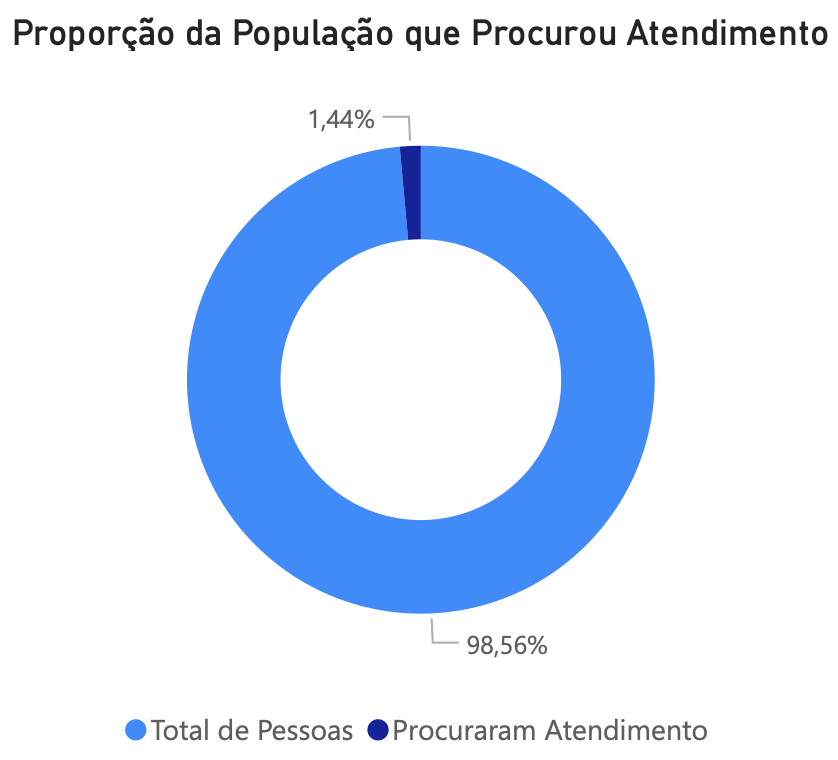
A análise da **distribuição geral de casos** reforça o cenário de **crescimento acelerado da pandemia** no período observado. Enquanto os **sintomas** mantiveram estabilidade entre maio e junho, há um salto expressivo em julho, acompanhado pelo aumento proporcional das **internações e entubações**. Esse comportamento indica uma **evolução progressiva da gravidade clínica**, sugerindo que o aumento dos sintomas foi seguido por uma **maior demanda hospitalar**, o que evidencia o avanço do contágio e o impacto sobre o sistema de saúde.

### **3.2 Distribuição territorial**

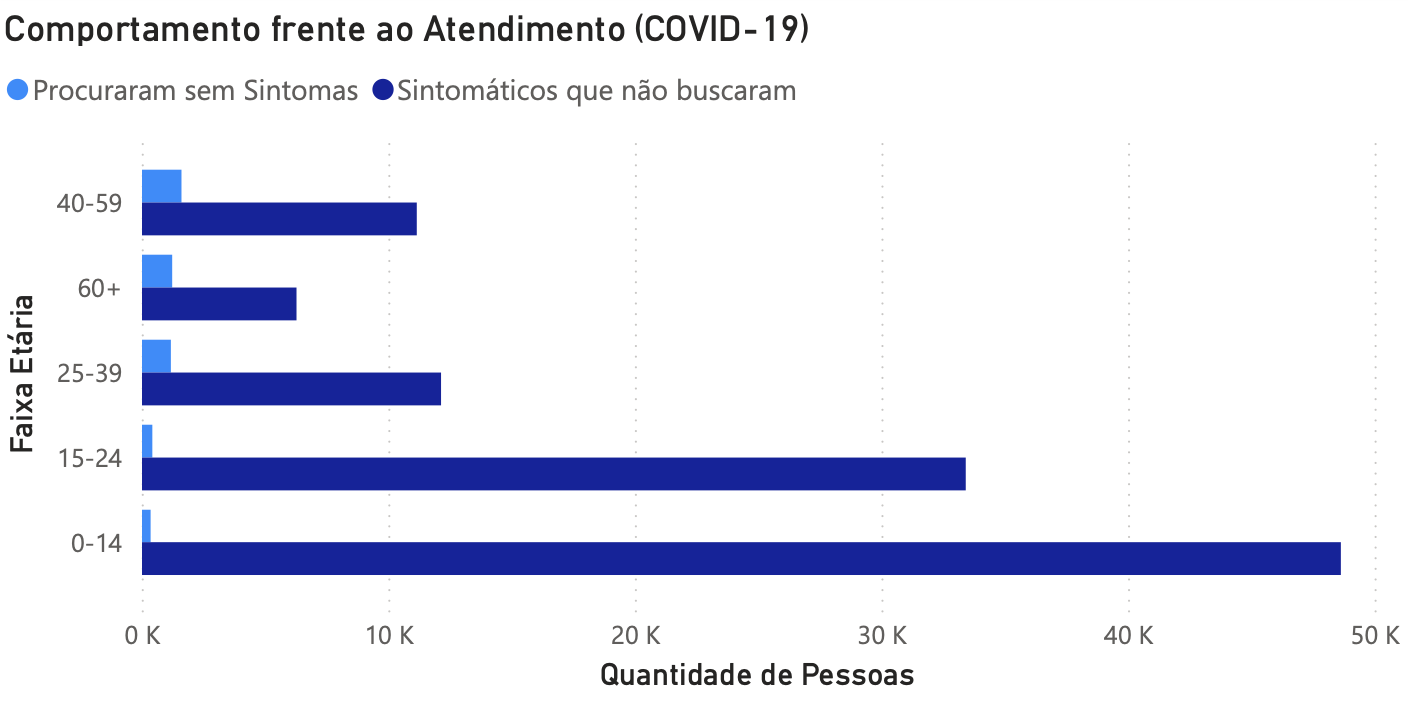


Os maiores volumes de notificações concentraram-se nos estados de **São Paulo**, **Minas Gerais**, **Maranhão** e **Rio de Janeiro**, o que reflete tanto a densidade populacional quanto as desigualdades regionais de infraestrutura hospitalar, conforme corroborado por análises nacionais.

### **3.3 Busca por atendimento**



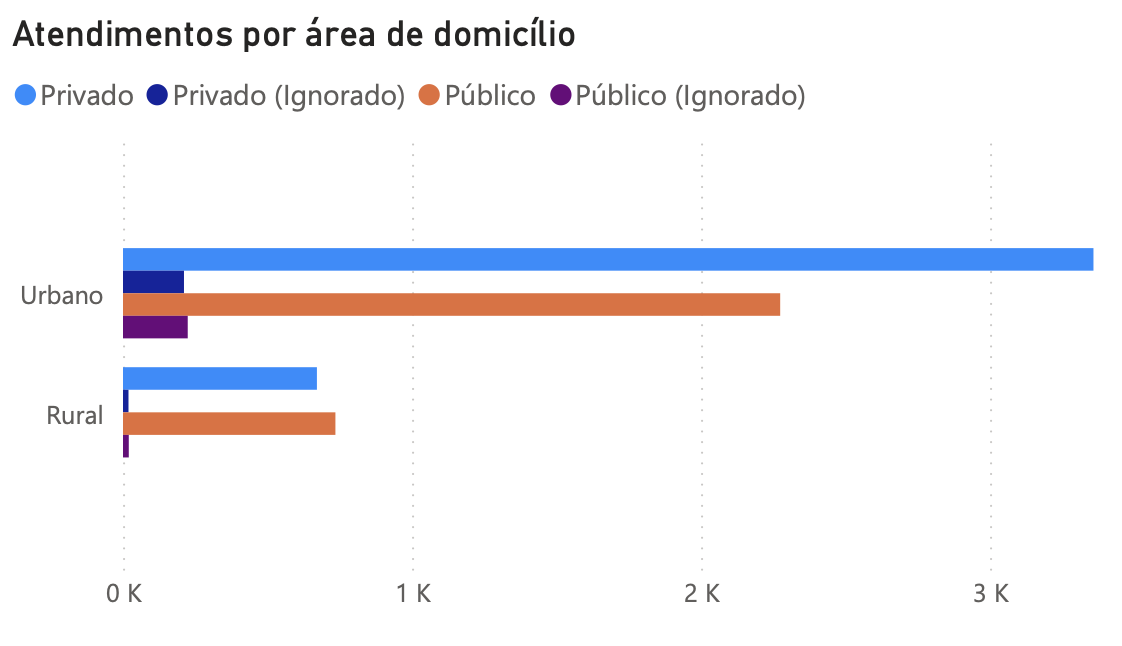
Apenas **1,44% da população pesquisada** declarou ter procurado atendimento médico, evidenciando uma **subutilização dos serviços de saúde** e, possivelmente, **subnotificação de casos leves**.

  
 Entre os que buscaram atendimento, observou-se predominância de jovens nas faixas de **15 a 24 anos** e **0 a 14 anos**. Por outro lado, os idosos apresentaram menor volume de procura, mas maior proporção de casos graves.



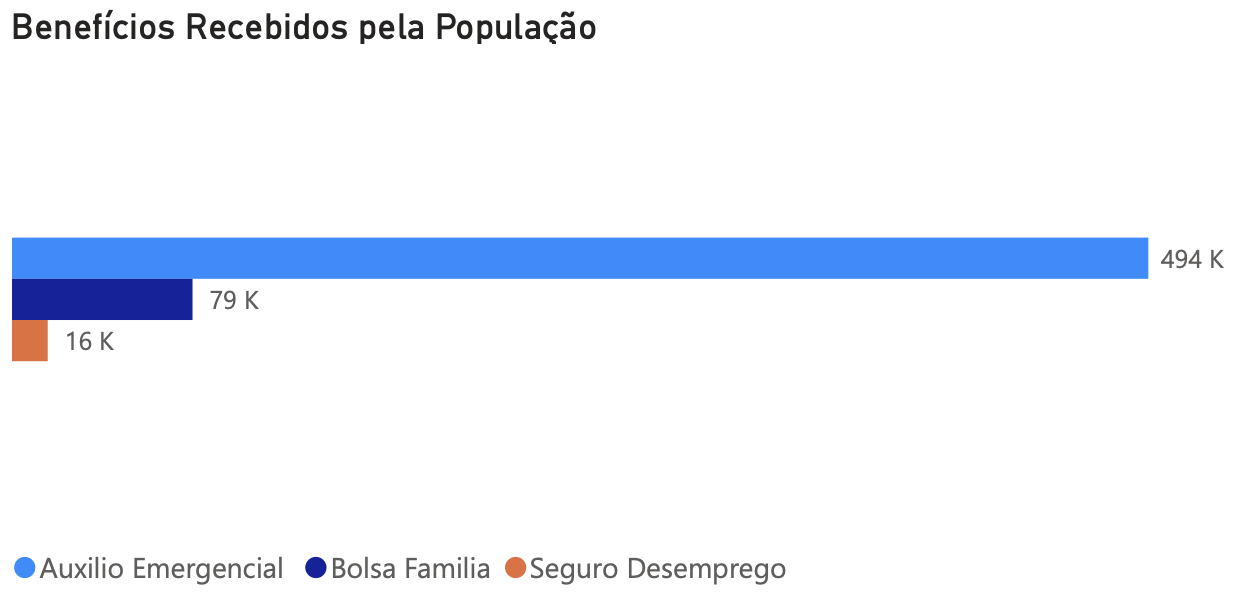
Em relação à escolaridade, as categorias **ensino médio completo** e **fundamental incompleto** apresentaram os maiores índices de busca, o que sugere relação entre **nível educacional e percepção de risco**.

### **3.4 Atendimento e área de residência**

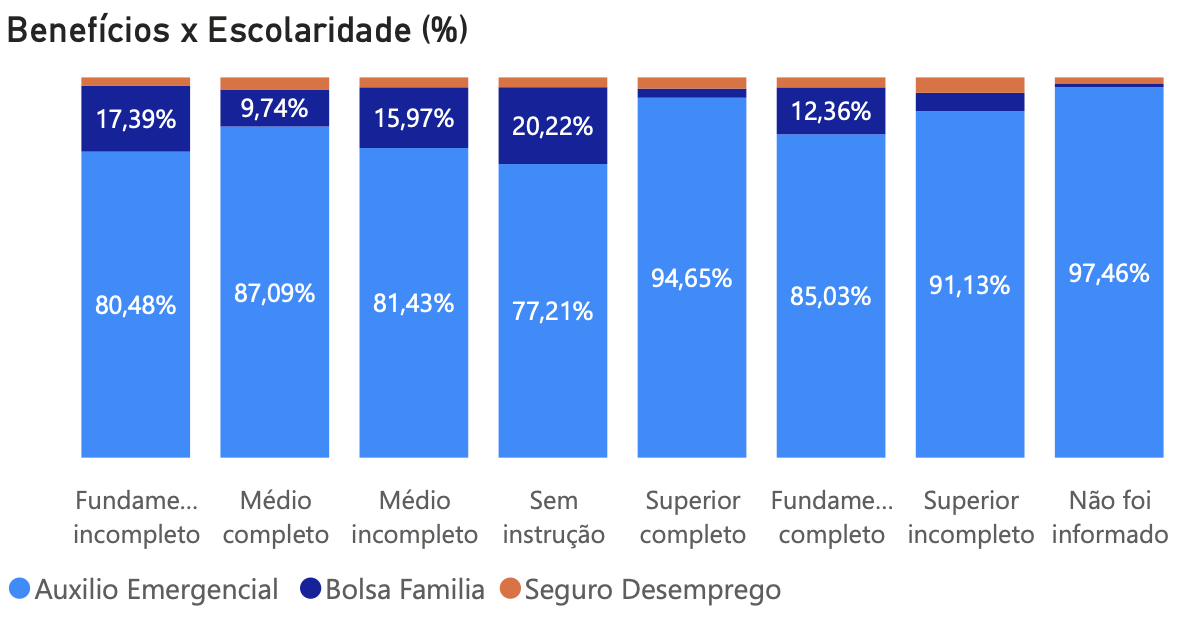


Mais de **80% dos atendimentos** foram registrados em áreas urbanas, com predominância do setor privado. Nas zonas rurais, o atendimento público foi majoritário, porém com volume inferior, indicando **limitações logísticas e carência de infraestrutura hospitalar**.  
Esse resultado reforça a influência das **desigualdades geográficas e econômicas** sobre o acesso ao tratamento e os desfechos clínicos, corroborando o padrão de disparidade identificado em análises de letalidade conduzidas pela PUC-Rio e FIOCRUZ.

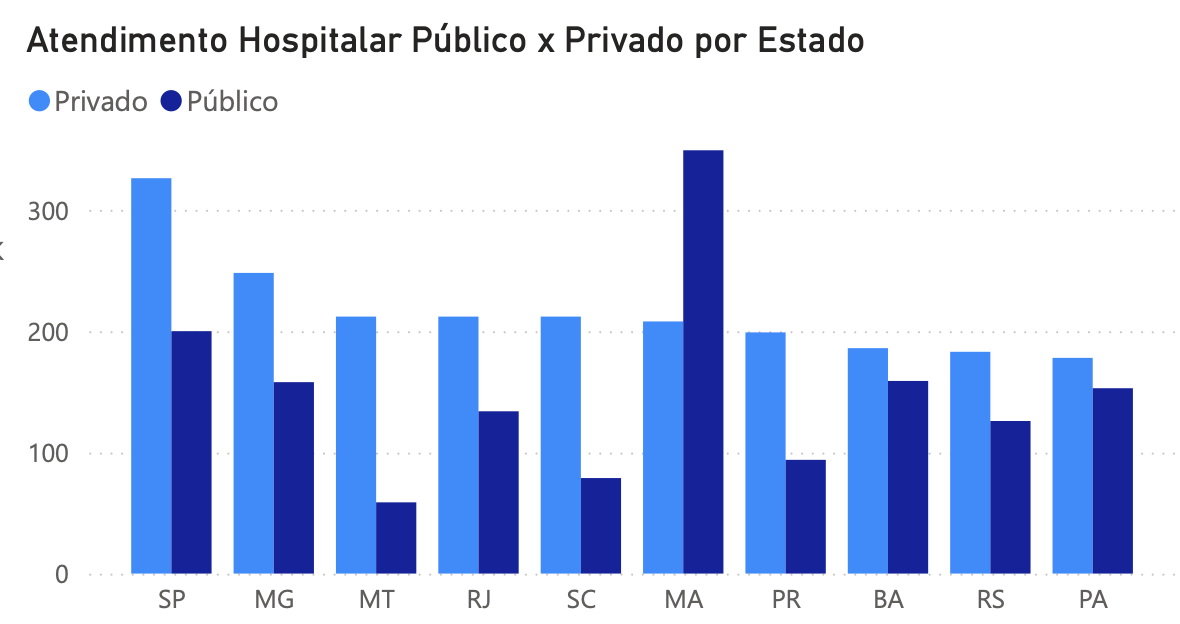
### **3.5 Dimensão socioeconômica**



O **Auxílio Emergencial** foi o benefício mais frequente, alcançando cerca de **494 mil pessoas**, seguido por **Bolsa Família** e **Seguro-Desemprego**.



Observa-se que o **Auxílio Emergencial** manteve predominância em todos os níveis de escolaridade, inclusive entre os indivíduos com **ensino superior completo (94,65%)** e **incompleto (91,13%)**, o que reforça o caráter universal da crise econômica durante a pandemia. Ainda assim, a **maior proporção de Bolsa Família** e **Seguro-Desemprego** aparece entre os grupos com **menor escolaridade** ou **sem instrução**, revelando que a vulnerabilidade social se concentra mais fortemente nessas faixas. Essa relação evidencia a correlação entre **baixa escolaridade e dependência de políticas de transferência de renda**, ao passo que o impacto da pandemia alcançou até mesmo camadas tradicionalmente mais estáveis do mercado de trabalho.



O gráfico mostra diferenças expressivas no acesso aos tipos de atendimento hospitalar entre os estados. Em locais como **Maranhão (MA)**, há predominância do **atendimento público**, indicando **menor oferta de serviços privados e maior dependência do SUS**. Já em **São Paulo (SP)** e **Minas Gerais (MG)**, o volume de atendimentos **privados supera os públicos**, refletindo uma estrutura hospitalar mais diversificada e com maior presença de planos de saúde. Essa variação regional demonstra as **desigualdades no sistema de saúde brasileiro**, onde o acesso a serviços privados está diretamente ligado ao **nível de desenvolvimento econômico e infraestrutura de cada estado**.  
 A distribuição dos benefícios demonstra que a crise sanitária impactou transversalmente diferentes grupos sociais, inclusive aqueles com maior escolaridade, revelando **fragilidade econômica generalizada** durante o período pandêmico.

## **4 Discussão**

Os resultados indicam que o comportamento da população brasileira durante a pandemia foi fortemente condicionado por fatores **socioeconômicos, territoriais e tecnológicos**.

1. **Fatores socioeconômicos:** a baixa renda e o baixo nível de escolaridade correlacionaram-se com menor probabilidade de busca por atendimento e maior risco de agravamento clínico.
2. **Fatores territoriais:** as desigualdades regionais impactaram o tempo de resposta e o acesso a recursos hospitalares, ampliando a vulnerabilidade de municípios de menor IDH.
3. **Fatores tecnológicos:** a necessidade de distanciamento social impulsionou a adoção de **tecnologias digitais em saúde**, como teleatendimento e monitoramento remoto, transformando a lógica assistencial.

Assim, a pandemia consolidou-se como um **marco da transformação tecnológica na saúde**, exigindo a incorporação definitiva de soluções digitais integradas à prática clínica. Ao mesmo tempo, evidenciou a urgência de políticas de **equidade territorial e social**, capazes de reduzir disparidades no acesso e na qualidade do cuidado.

## **5 Recomendações estratégicas**

|  |  |  |
| --- | --- | --- |
| **Eixo** | **Ação proposta** | **Justificativa** |
| **1. Planejamento hospitalar** | Dimensionar leitos e insumos com base na proporção de **2,2% de casos graves** sobre sintomáticos. | Antecipar a saturação e otimizar recursos. |
| **2. Vigilância ativa** | Monitorar sintomas respiratórios via teleatendimento e plataformas digitais. | Detecção precoce de agravamentos. |
| **3. Saúde digital** | Ampliar **teleconsultas, monitoramento remoto e inteligência artificial clínica**. | Garantir continuidade de cuidado com menor exposição. |
| **4. Acesso equitativo** | Implantar unidades móveis e parcerias com atenção básica em regiões periféricas. | Corrigir desigualdades regionais. |
| **5. Comunicação segmentada** | Campanhas digitais voltadas a **jovens e trabalhadores informais**. | Reduzir subnotificação e ampliar adesão. |
| **6. Integração social** | Parcerias com órgãos de assistência e transporte sanitário. | Apoiar populações vulneráveis. |
| **7. Inteligência de dados** | Criar **painel dinâmico** com métricas de sintomas, gravidade e internações. | Apoiar decisões estratégicas em tempo real. |

## **6 Conclusão**

A análise descritiva da PNAD-COVID-19 demonstra que o impacto da pandemia ultrapassou o campo epidemiológico, evidenciando a interdependência entre **determinantes sociais, acesso à saúde e infraestrutura tecnológica**.  
 A resposta hospitalar deve, portanto, basear-se em **planejamento orientado por dados**, **integração intersetorial** e **adoção de tecnologias de cuidado remoto**, com o objetivo de mitigar desigualdades e garantir atendimento universal e eficaz em futuros surtos.

A experiência da COVID-19 reforça que a eficiência do sistema de saúde depende não apenas de sua capacidade assistencial, mas também da **capacidade analítica e adaptativa** frente às transformações tecnológicas e sociais.

## **Referência**

NÚCLEO DE OPERAÇÕES E INTELIGÊNCIA EM SAÚDE (NOIS). *Análise socioeconômica da letalidade dos casos da COVID-19 no Brasil.* Nota Técnica nº 11. Rio de Janeiro: PUC-Rio, Instituto D’Or de Pesquisa e Ensino, 2020. Disponível em:<https://sites.google.com/view/nois-pucrio>. Acesso em: 05 out. 2025

Organização Pan-Americana da Saúde. Folha informativa - COVID-19 (doença causada pelo novo coronavírus).

Ministério da Saúde declara transmissão comunitária nacional. G1 2020; 20 mar.

Ministério da Saúde. Coronavírus: o que você precisa saber e como prevenir o contágio. https://coronavirus.saude.gov.br (acessado em 09/Jul/2020).