Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciência
Instituto de Matemática e Estatística

Gabriel Pereira Mendes

Uma Proposta para Análise da Prevalência de Alelos HLA em Pacientes com COVID-19

Rio de Janeiro
2023

Gabriel Pereira Mendes

Uma Proposta para Análise da Prevalência de Alelos HLA em Pacientes com COVID-19

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Alexandre da Costa Sena
Orientador: Prof. Dr. Luís Cristóvão de Moraes S. Pôrto

Rio de Janeiro
2023

CATALOGAÇÃO NA FONTE

UERJ/REDE SIRIUS/BIBLIOTECA CTC/A

M538 Mendes, Gabriel Pereira.
Uma proposta para análise da prevalência de alelos HLA em pacientes com Covid -19 / Gabriel Pereira Mendes. – 2023.
55 f.: il.

Orientadores: Alexandre da Costa Sena, Luís Cristovão de Moraes S. Pôrto
Dissertação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística.

1. Algoritmos de computador - Teses. 2. Antígenos HLA - Teses. 3. COVID-19 (Doença) - Teses. 4. Saúde - Processamento de dados. I. Sena, Alexandre da Costa. II. Pôrto, Luís Cristovão de Moraes S. III. Universidade do Estado do Rio de Janeiro. Instituto de Matemática e Estatística. IV. Título.

CDU 004.421

Patricia Bello Meijinhos CRB7/5217 - Bibliotecária responsável pela elaboração da ficha catalográfica

_____ _____

Assinatura Data

Gabriel Pereira Mendes

Uma Proposta para Análise da Prevalência de Alelos HLA em Pacientes com COVID-19

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Aprovada em 13 de Março de 2023.
Banca Examinadora:

Prof. Dr. Alexandre da Costa Sena (Orientador)
Instituto de Matemática e Estatística – UERJ

Prof. Dr. Luís Cristóvão de Moraes S. Pôrto (Orientador)
Professor Titular Instituto de Biologia Roberto Alcantara Gomes - UERJ

Prof.ª Dra. Karla Tereza Figueiredo Leite
Professora Adjunta Instituto de Matemática e Estatística - UERJ

Dra. Juliana Fernandes Cardoso
CareDx, Inc.

Rio de Janeiro
2023

DEDICATÓRIA

Dedico essa obra a Deus, a minha mãe, a alguns outros familiares e também amigos que me apoiaram durante todo o meu mestrado.

AGRADECIMENTOS

Agradeço a minha família, amigos e aos meus orientados na elaboração deste trabalho. Gratificado por todo o apoio durante a escrita dessa produção acadêmica. Esse projeto foi desenvolvido pensando-se em ajudar a comunidade científica de uma maneira geral e espera-se que ele possa ser usado como fonte de referência para futuros trabalhos. Além disso, tem-se uma expectativa de que o tema de estudo aqui desenvolvido se converta em benefícios para a sociedade no contexto científico e tecnológico no qual ele está inserido.

A sabedoria comunica a vida a seus filhos
e acolhe os que a procuram.
Bíblia Sagrada. Eclo 4,12

RESUMO

MENDES, Gabriel Pereira. *Uma Proposta para Análise da Prevalência de Alelos HLA em Pacientes com COVID-19*. 2023. 55 f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística
, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

Os sistemas de informação SIVEP-Gripe e e-SUS são utilizados para registrar notificações de casos de COVID-19. Enquanto o SIVEP-Gripe é um sistema de vigilância para registrar casos generalizados de Síndrome Respiratória Aguda Grave (SRAG), que inclui também casos graves de infecções ou óbitos causados pelo SARS-CoV-2, o sistema e-SUS foi criado durante a pandemia para registrar casos suspeitos ou confirmados de COVID-19. Além disso, o Ministério da Saúde do Brasil também conta com base de dados com as características genéticas de doadores de células tronco hematopoéticas (REDOME), onde a compatibilidade desses genes, além ter um papel fundamental no processo de rejeição, está diretamente ligada a resposta imunológica para patógenos (vírus, bactérias e protozoários). Assim, a partir dos registros dessas bases de dados, o objetivo central deste trabalho é, em primeiro lugar, gerar uma base de dados confiável para um Estudo Caso-Controle (ECC) da frequência de alelos HLA em pacientes com COVID-19, e, em seguida, realizar a análise da frequência de alelos HLA na base SIVEP-Gripe em relação a base e-SUS. Este trabalho apresenta todas as etapas necessárias para atingir esse objetivo e, em especial, propõe, implementa e avalia um algoritmo de pareamento para criar uma base de controle homogênea, a partir das características dos registros da base de casos. Por fim, o trabalho apresenta a prevalência de alelos HLA para pacientes com COVID-19 do estado de Minas Gerais. Os resultados mostram que o algoritmo proposto é capaz de gerar bases comparáveis em número de casos e controle proporcionais para cinco variáveis (i.e. local, idade, etnia, período de testagem e tipo de teste) e que existem diferenças na distribuição dos grupos alélicos em função do campo raça/etnia, onde o alelo B*51 tem uma chance maior de oferecer proteção, enquanto que o alelo A*36 aumenta o risco para a COVID-19.

Palavras-chave: COVID-19. HLA. Algoritmo de Pareamento.

ABSTRACT

MENDES, Gabriel Pereira. *A Proposal for Analysis of the Prevalence of HLA Alleles in Patients with Covid-19*. 2023. 55 f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística
, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

The SIVEP-Gripe and e-SUS information systems are used to record COVID-19 notifications. While SIVEP-Gripe is a surveillance system to record cases of Severe Acute Respiratory Syndrome (SARS) that also include serious cases of infections or deaths caused by SARS-CoV-2, the e-SUS system was created during the pandemic to register suspected or confirmed cases of COVID-19. In addition, the Brazilian Ministry of Health also has a database with the genetic characteristics of hematopoietic stem cell donors (REDOME), where the compatibility of these genes, in addition to playing a fundamental role in the rejection process, is directly linked to the immune responses to pathogens (viruses, bacteria and protozoa). Therefore, based on the records from these databases, the main objective of this work is, firstly, to generate a reliable database for an Case-Control Study (CCS) of the prevalence of HLA alleles in patients with COVID-19, and then to carry out the analysis of the frequency of HLA alleles. This paper presents the necessary steps to achieve this objective and, in particular, proposes, implements and evaluates a matching algorithm to create a homogeneous control base, based on the characteristics of the case base records. Finally, this work presents the prevalence of HLA alleles for patients with COVID-19 in the state of Minas Gerais. Results show that the proposed algorithm is able to generate comparable bases in number of cases and control proportional for five variables (i.e. location, age, ethnicity, testing period and test type) and that there are differences in the distribution of alleles groups according to the race/ethnicity field, where the B*51 allele has a greater chance of offer protection, while the A*36 allele increases the risk for COVID-19.

Keywords: COVID-19. HLA. Matching Algorithm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SUS	Sistema Único de Saúde
COVID-19	Coronavírus
SIS	Sistemas de Informação em Saúde
SIVEP-Gripe	Sistema de Informação da Vigilância Epidemiológica da Gripe
e-SUS	e-SUS Notifica
SG	Síndrome Gripal
SRAG	Síndrome Respiratória Aguda Grave
DATASUS	Departamento de Informática do Sistema Único de Saúde
FUNED	Fundação Ezequiel Dias
HLA	Human Leukocyte Antigen
REDOME	Registro Nacional de Doadores Voluntários de Medula Óssea
REREME	Registro Nacional de Receptores de Medula Óssea
SES	Secretaria Estadual de Saúde
SMS	Secretaria Municipal de Saúde
CIT	Comissão Intergestores Tripartite
CIB	Comissão Intergestores Bipartite
CONASS	Conselho Nacional de Secretário da Saúde
CONASEMS	Conselho Nacional de Secretarias Municipais de Saúde
COSEMS	Conselhos de Secretarias Municipais de Saúde
H1N1	Influenza A
CSV	Comma-separated values
JPA	Java Persistence API
ORM	Object-Relational Mapping
ECC	Estudo Caso-Controle
LGPD	Lei Geral de Proteção de Dados
UTI	Unidade de Terapia Intensiva
UPA	Unidade de Pronto Atendimento
CDC	Centers for Disease Control and Prevention
CONEP	Comissão Nacional de Ética em Pesquisa

SUMÁRIO

INTRODUÇÃO

Políticas voltadas para a *gestão de saúde pública* são algumas das principais preocupações dos governos pois, elas afetam, direta ou indiretamente, todos os setores de uma nação como a ciência, a economia, a política, a educação, os meios de transporte, entre outros. Mais especificamente, pode-se pensar na saúde como um setor particular da economia responsável pela geração de *bens* (atendimentos, consultas ou exames) e *serviços* (vacinas ou medicamentos) que visam zelar pelo *bem-estar* das pessoas (Paim, Jairnilson Silva, 2009).

Assim como em outras áreas, a saúde foi um dos campos que passou por uma *informatização* de seus processos incluindo também o SUS. Grande parte das tarefas que antes eram executadas manualmente por profissionais e unidades de saúde foram automatizadas e delegadas para os chamados SIS (Sistemas de Informação em Saúde). Estes sistemas ficaram responsáveis pelo apoio no *planejamento estratégico de ações* (de Fátima Marin, Heimar, 2010; Cavalcante, Ricardo Bezerra and Ferreira, Marina Nagata and Silva, Poliana Cavalcante, 2011) voltadas para a gestão de saúde pública.

O controle epidemiológico da COVID-19 (Santos, Alethele de Oliveira and Lopes, Luciana Tolêdo, 2021; Freitas, Carlos Machado de and Barcellos, Christovam and Villela, Daniel Antunes Maciel, 2021) foi um dos problemas sanitários no qual se evidenciou a importância de se integrar eficientemente os sistemas de informação usados nos processos do SUS. Durante o combate à pandemia no Brasil, dois destes sistemas de informação se destacaram para a gestão de saúde pública (Brasil, Ministério da Sáude, b): o SIVEP-Gripe e o e-SUS. Ambos são usados para registrar *notificações* e apoiar no processo de *monitoramento* de casos de COVID-19 (Freitas, Carlos Machado de and Barcellos, Christovam and Villela, Daniel Antunes Maciel, 2021).

O SIVEP-Gripe (Sistema de Informação da Vigilância Epidemiológica da Gripe) é um sistema de vigilância que já era usado para notificar casos generalizados de SRAG (Síndrome Respiratória Aguda Grave) mas, passou também a ser usado para registrar casos graves de infecções ou óbitos causados especificamente pelo SARS-CoV-2. Devido a alta demanda de notificações, surgiram também outros sistemas para apoiar neste processo de monitoramento de infecções, como o e-SUS, usado para registrar casos de SG (Síndrome Gripal), suspeitos ou confirmados, de COVID-19 (Brasil, Ministério da Sáude, b).

Por outro lado, o Ministério da Saúde do Brasil também conta com base de dados com as características genéticas que permitem a busca de possíveis doadores de células tronco hematopoéticas (REDOME - Registro Nacional de Doadores Voluntários de Medula Óssea) para pacientes que necessitam receber essas células como tratamento (REREME-Registro Nacional de Receptores de Medula Óssea) (Brasil, Ministério da Sáude, f). As compatibilidades desses genes além de serem definidoras do processo de rejeição estão

implicadas nas respostas imunológicas para patógenos (vírus, bactérias e protozoários) e dependendo da combinação de um ou mais alelos desses genes estão associados com a susceptibilidade de determinadas doenças e uma resposta imunológica efetiva após a vacinação.

Assim, o objetivo central deste trabalho é propor e implementar uma abordagem para criação de uma base de dados confiável para avaliação da *frequência de alelos* HLA (Human Leukocyte Antigen) em pacientes com COVID-19, a partir dos dados das bases SIVEP-Gripe e e-SUS. Em seguida, com a base gerada, avaliar a prevalência de alelos HLA em pacientes que tiveram COVID-19. Uma base de dados confiável com dados sobre pacientes que tiveram COVID-19 e, também, informações sobre os seus alelos HLA podem auxiliar aos especialistas em imunologia a identificar não somente alelos que ajudem na proteção contra a COVID-19, mas também alelos que indiquem um maior risco de ter complicações (Correale P, Mutti L, Pentimalli F, et al., 2020).

Para atingir esse objetivo, as seguintes etapas foram necessárias: *pré-processamento* que incluiu *uniformização*, *limpeza*, e *integração* das bases de dados; *busca* dos dados sobre os alelos na base de dados do REDOME; *criação* de uma base ECC (Estudo Caso-Controle) (S KERNDT CC, 2022) confiável. Todas essas etapas são descritas neste trabalho, em especial, o algoritmo de *pareamento* escalonável em até 7 (sete) variáveis e dimensionável na relação caso:controle - 1:N (e.g. 1:2, 1:3, 1:4, etc.) para criação de uma base para ECC. Esse algoritmo tem potencial para ajudar bastante a comunidade científica, não só permitindo a criação de uma base de controle balanceada em relação aos casos baseando-se em características identificadas como *fatores de exposição ao risco* para a COVID-19, mas também por evitar a necessidade do pesquisador ter que escolher manualmente os casos da base de controle, evitando assim um possível *viés de seleção* que é um dos problemas da técnica de ECC (S KERNDT CC, 2022).

Tal algoritmo de pareamento é responsável por selecionar para cada registro da base de casos um ou mais registros da base de controle elaborando assim um ECC (STRALEN et al., 2010). Em um ECC busca-se a *relação entre uma doença com um fator de exposição ao risco* verificando-se a distribuição da frequência desse fator nos casos e nos controles. Por exemplo, em (OLIVEIRA; VELLARDE; SÁ, 2015), é abordado se o consumo de álcool (fator de risco) aumenta as chances de surgimento de câncer de pâncreas (doença). No escopo deste trabalho, a doença do ECC aqui elaborado é a COVID-19 sobre a qual deseja-se avaliar se existe uma prevalência de alelos HLA que ajudem na proteção contra COVID-19 e, também, se existe prevalência de alelos que indique um maior risco de ter complicações (Correale P, Mutti L, Pentimalli F, et al., 2020). Assim, os casos serão os pacientes testados positivo para COVID-19 da base SIVEP-Gripe, uma vez que esta base é composta de casos graves de SARS-Cov-2. Por sua vez, os controles serão escolhidos entre os pacientes positivos e negativos para COVID-19 da base e-SUS, que é composta de casos suspeitos ou confirmados de COVID-19 leve.

O balanceamento entre a base de controle e a base de casos para a avaliação da prevalência de alelos HLA é muito importante, especialmente quando se considera a etnia dos pacientes que tem uma influência direta no alelo que pode ser encontrado, ainda mais quando se considera populações miscigenadas, que traçam sua ascendência através de diferentes regiões geográficas. Por exemplo, o trabalho (NUNES et al., 2020) mostrou que pacientes com ascendência africana enfrentam maior dificuldade em localizar potenciais doadores em função da maior variabilidade HLA das populações Africanas e da maior presença de doadores brancos no banco de dados de doadores de órgãos avaliado no estudo. Por outro lado, o ECC para a prevalência de alelos pode ajudar a identificar alelos que possam ajudar a fornecer uma proteção ou um risco maior a COVID-19. Por exemplo, no estudo (CORREALE et al., 2020) foram identificados alelos relacionados a disseminação da COVID-19 na Itália.

A avaliação do algoritmo de pareamento proposto, utilizando duas bases de dados distintas (RJ e MG), mostrou que ele é capaz de produzir um arquivo de controle *homogêneo*, especialmente para os campos de maior prioridade. Além disso, ao se comparar com outros dois algoritmos (**Sem Filtro** e **Aleatório**) que não utilizam nenhuma prioridade, é possível verificar a eficiência do algoritmo proposto. Por fim, a análise da prevalência de alelos HLA para os pacientes do estado de Minas Gerais mostrou que existem diferenças na distribuição dos grupos alélicos em função do campo raça/etnia (MENDES et al., 2022).

Este trabalho está divido da seguinte forma. O Capítulo 1 apresenta a fundamentação teórica, mostrando aspectos essenciais para o entendimento da motivação do trabalho. Por sua vez, no Capítulo 2, é apresentada a metodologia adotada neste trabalho. Em seguida, no Capítulo 3, são apresentados os resultados sobre a avaliação do algoritmo de pareamento e prevalência de alelos HLA em pacientes com COVID-19. Por fim, apresentam-se as conclusões e perspectivas de trabalhos futuros.

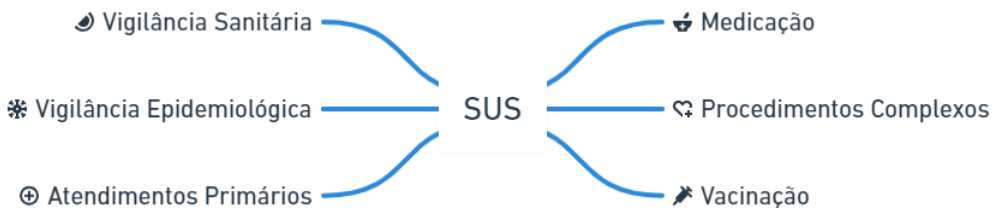
1 REVISÃO DA LITERATURA

Neste capítulo, serão apresentados com mais detalhes as leis que regem o SUS e o papel que esse sistema exerceu na reforma sanitária brasileira. Além disso, é apresentado um pouco mais de informações sobre os sistemas e-SUS e SIVEP-Gripe especificando melhor os seus objetivos pretendidos e funcionalidades oferecidas. Por fim, este capítulo apresenta conceitos importantes sobre ECC e alguns trabalhos relacionados.

1.1 SUS (Sistema Único de Saúde)

O SUS (Paim, Jairnilson Silva, 2009; Paim, Jairnilson Silva, 2018) é o sistema público de saúde brasileiro, que foi instituído pela Constituição Federal de 1988 (Federal, Senado, 1988) no seu artigo 196, sendo este um dos maiores sistemas de saúde do mundo atendendo mais de 190 milhões de habitantes. Conforme ilustrado pela Figura 1, o SUS engloba vários tipos de *serviços* que são realizados diariamente por profissionais e órgãos de saúde. Os serviços incluem os mais comuns entre a população como *consultas*, *exames* e *atendimentos ambulatoriais*, além de procedimentos mais complexos como *cirurgias* ou *transplantes de órgãos*. Além disso, o SUS também é responsável por regular a *vigilância sanitária* de alimentos e a *vigilância epidemiológica*, ou seja, o controle de epidemias ou pandemias. Também é responsabilidade do SUS prover recursos para a vacinação e a medicação da população no combate à doenças infecciosas e transmissíveis. Sendo um sistema público, seu financiamento vem basicamente dos impostos cobrados da população e colhidos pelas três esferas do governo.

Figura 1 - SUS (Sistema Único de Saúde).



Antes da criação do SUS, a saúde não era considerada um *direito social* e para ter acesso a ela, os indivíduos necessitavam pagar pelos serviços ou serem contribuintes da previdência social. Naquela época, o indivíduo precisava ter recursos financeiros ou algum tipo de mérito para conseguir acessar um serviço de saúde ou dependia da caridade alheia oferecida, por exemplo, por algumas instituições filantrópicas existentes.

Com o SUS, instaurou-se o acesso gratuito a serviços de saúde de forma integral e

universal e, como a maior parte da população brasileira é dependente quase que exclusivamente deste sistema para o cuidado de seu bem estar físico ou mental, este acontecimento foi um marco importante para a *reforma sanitária* do país. Em resumo, o SUS *democratizou* e *universalizou* o acesso à saúde.

Além disso, a criação do SUS também marcou uma mudança no conceito de saúde que antes era apenas caracterizada como *voltado para o tratamento de doenças* e agora passa a englobar também a *prevenção de enfermidades*. Desde então, tanto o tratamento quanto a prevenção passaram a compor o planejamento de ações adotadas pelas políticas de saúde pública.

Desde a sua criação, o SUS passou por vários processos de informatização dos seus serviços. Vários sistemas de informação (os chamados SIS) foram criados para facilitar a integração e coordenação das ações voltadas à política de saúde pública (PINTO; FREITAS; FIGUEIREDO, 2018; Brasil, Ministério da Sáude, c; Brasil, Ministério da Sáude, e; Brasil, Ministério da Sáude, d).

1.1.1 Sistemas de Informação do SUS e pandemia da COVID-19

A pandemia causada pela COVID-19 motivou o uso de sistemas de informação (já existentes ou criados) para auxiliar no monitoramento de infecções causadas pelo vírus SARS-CoV-2 (Freitas, Carlos Machado de and Barcellos, Christovam and Villela, Daniel Antunes Maciel, 2021). Para o combate à pandemia no Brasil, o SUS teve o auxílio de sistemas de informação, sendo que dois deles se destacaram pelo seu alto grau de importância para a notificação de casos de COVID-19:

- O SIVEP-Gripe usado para notificar casos mais graves ou de óbitos pela doença;
- O e-SUS usado para notificar casos suspeitos ou confirmados (leves a moderados) da doença.

Esses sistemas são alimentados com dados fornecidos por profissionais ou unidades de saúde vinculados ao SUS e são usados pelos governos, órgãos e instituições no planejamento de ações voltadas para o combate à pandemia.

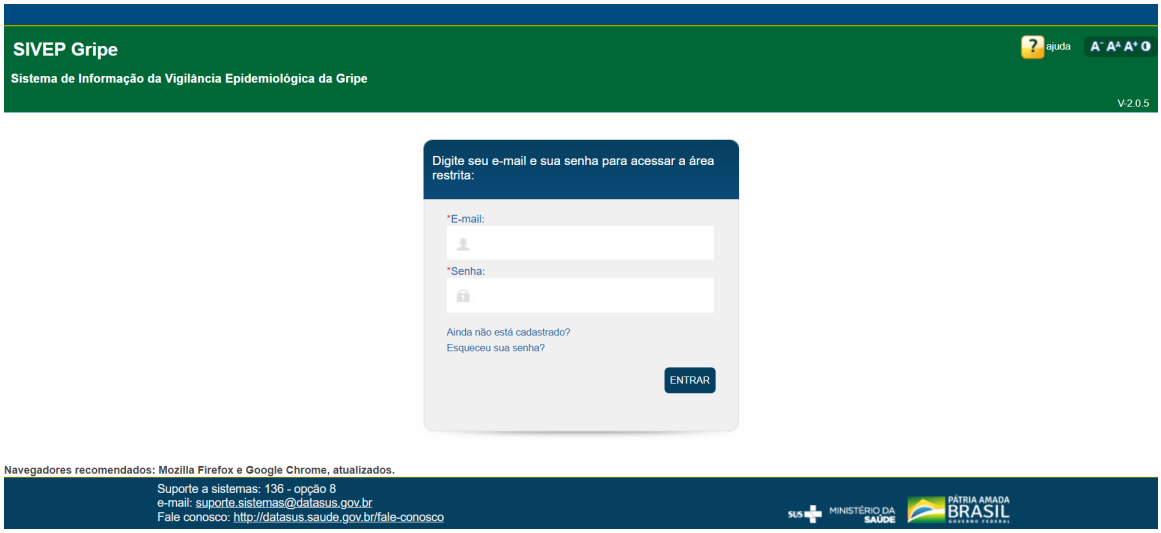
1.1.2 SIVEP-Gripe (Sistema de Informação da Vigilância Epidemiológica da Gripe)

O SIVEP-Gripe (Brasil, Ministério da Sáude, g) atua desde o ano 2000 e foi originalmente criado para monitorar casos de SG (Síndrome Gripal) causadas pelo vírus influenza. Alguns marcos marcaram atribuições importantes para este sistema (Bahia, Governo do Estado da,):

- em 2009, com a pandemia causada pelo vírus H1N1 (Influenza A), foi fortalecido o controle epidemiológico de vírus respiratórios e instaurou-se a vigilância da SRAG (Síndrome Respiratória Aguda Grave);
- em 2020, com a pandemia causada pela COVID-19, o SIVEP-Gripe passou também a notificar casos de SRAG causadas pelo SARS-CoV-2.

A Figura 2 ilustra a página de login [1] do SIVEP-Gripe. O usuário precisa se cadastrar para utilizar a ferramenta fornecendo dados pessoais além do seu número do cartão nacional de saúde que pode ser obtido em uma unidade de saúde. Na página também encontram-se informações de contato dos responsáveis pelo suporte ao sistema.

Figura 2 - Página de login SIVEP-Gripe.



Existe também uma outra página [2] em que é possível visualizar ou exportar dados sobre notificações de SRAG, referentes ao estado do Rio de Janeiro. É possível aplicar diversos filtros além de poder ver os dados em tabelas ou gráficos como os de dispersão, linhas, colunas ou setores (Figura 3).

Por último, conforme ilustrado na Figura 4, existe o site INFOGripe [3] no qual é possível visualizar gráficos com dados sobre notificações registradas no SIVEP-Gripe. Os dados indicam para cada estado, capital ou macrorregião, como está a curva atual de contágio da doença (se está em tendência de crescimento, queda ou estabilidade) além de mostrar o número de casos semanais juntamente com a média móvel.

[1] ⟨https://sivepgripe.saude.gov.br/sivepgripe/login.html⟩

[2] ⟨http://sistemas.saude.rj.gov.br/tabnetbd/dhx.exe?sivep_gripe/sivep_gripe.def⟩

[3] ⟨http://info.gripe.fiocruz.br/⟩

Figura 3 - Página de notificações sobre SIVEP-Gripe RJ.

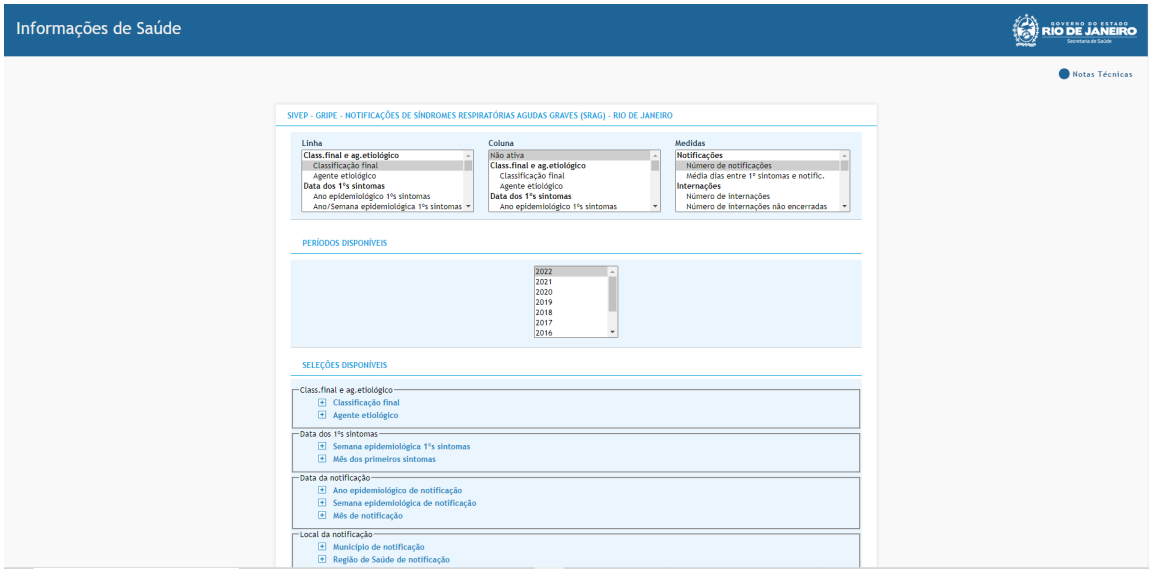
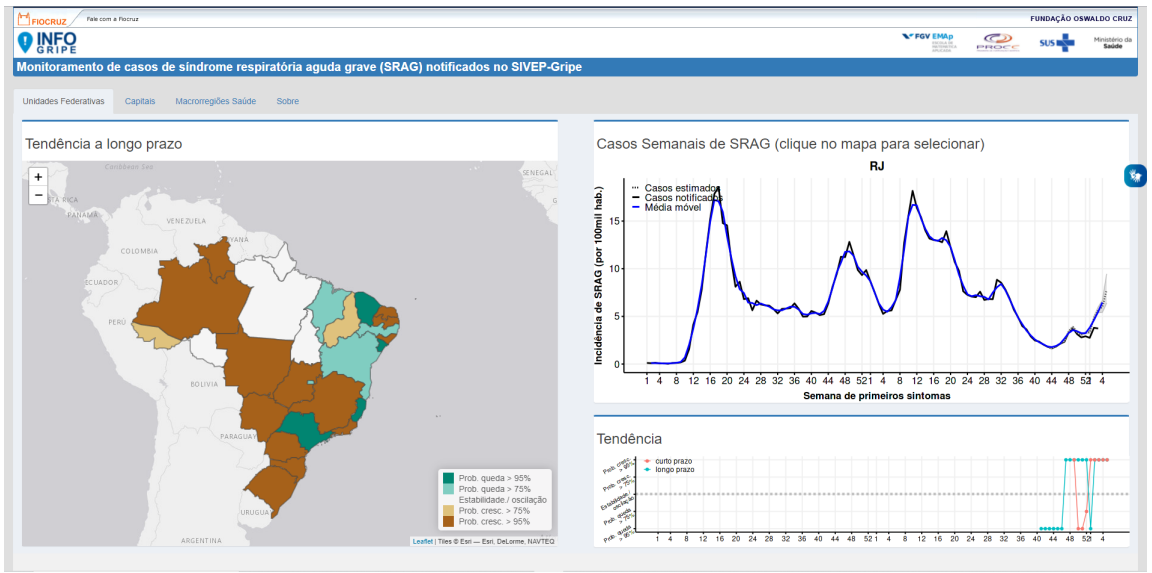


Figura 4 - Página do INFOGripe.

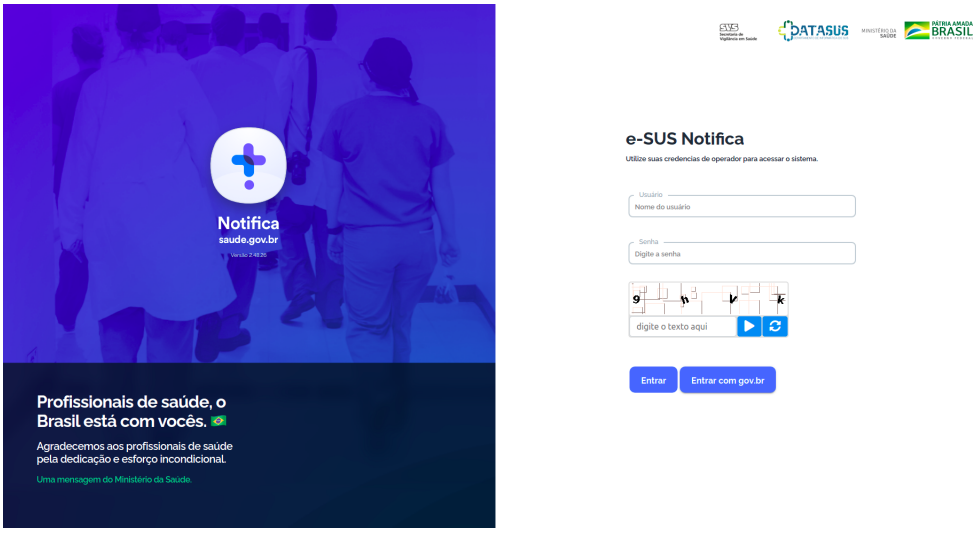


1.1.3 e-SUS (e-SUS Notifica)

No início da pandemia causada pela COVID-19 (Santos, Alethele de Oliveira and Lopes, Luciana Tolêdo, 2021; Freitas, Carlos Machado de and Barcellos, Christovam and Villela, Daniel Antunes Maciel, 2021), foi criado o e-SUS (Brasil, Ministério da Sáude, a) visando auxiliar no combate e no controle epidemiológico da doença. Nele são registrados os casos de SG suspeita ou confirmada de COVID-19 e todas estas notificações ficam disponíveis para serem consultadas de maneira ágil pelo sistema. Criado e mantido pelo DATASUS (Departamento de Informática do Sistema Único de Saúde) (Brasil, Ministério da Sáude, h), o e-SUS provê acesso às notificações, em tempo real, pelos profissionais e unidades de saúde.

Conforme ilustrado na Figura 5, o sistema contém uma página de login [4]. Similar ao SIVEP-Gripe, para utilizar o e-SUS é necessário que o usuário se cadastre na ferramenta fornecendo seus dados pessoais. Dentre as funcionalidades mencionadas pelo tutorial de navegação (Saúde, Ministério da,) é possível adicionar, atualizar, visualizar, cancelar, imprimir e exportar notificações além de poder consultar o histórico das mesmas. Usuários com perfil de gestor municipal ou estadual também podem acessar um módulo que permite gerenciar os dados de outros usuários do sistema.

Figura 5 - Página do e-SUS (e-SUS Notifica).



O sistema já passou por diversos aperfeiçoamentos para garantir melhor desempenho e atualmente conta com alguns módulos sendo os principais:

- *Notificação COVID-19*: responsável por auxiliar no controle epidemiológico da

[4] ⟨https://notifica.saude.gov.br/login⟩

COVID-19 registrando casos de SG suspeita ou confirmada de COVID-19 e também todos os resultados dos testes realizados para detecção da doença sejam eles positivos ou negativos;

- *Monitoramento de contatos*: responsável por monitorar indivíduos que tiveram contato com caso suspeito ou confirmado de COVID-19, sendo estes registrados no módulo *Notificação Covid-19*.

1.2 ECC (Estudo Caso-Controle)

O algoritmo proposto tem a motivação de gerar dados já balanceados entre caso-controle e sem a influência de qualquer *viés* (em inglês *biased*) (RÊGO, 2010). Em um ECC , os vieses são fatores que atrapalham na análise dos dados, sendo alguns deles o *viés de seleção* e o *viés de memória*. O viés de seleção pode ocorrer, por exemplo, quando usa-se maior quantidade de casos *prevalentes* (de longa duração) do que de casos *incidentes* (novos), deixando os resultados mais propensos a estarem distorcidos (RÊGO, 2010).

Em suma, com a preferência por casos incidentes, todos os indivíduos terão a mesma probabilidade de serem selecionados independente da evolução da doença. Contudo, com uma abordagem baseada em casos prevalentes, apenas os casos mais leves ou moderados serão passíveis de serem selecionados já que é possível que os casos mais graves da doença venham a óbito mais precocemente. Assim, apenas os casos sobreviventes ou prevalentes, que apresentam grau mais leve ou moderado da doença, poderiam ser selecionados levando-se a um viés de seleção. Por sua vez, segundo (OLIVEIRA; VELLARDE; SÁ, 2015), o viés de memória ocorre quando há uma divergência entre os casos e os controles quanto à recordação da exposição ao fator de risco.

Assim como descrito em (OLIVEIRA; VELLARDE; SÁ, 2015), a ideia central do algoritmo de pareamento deste trabalho é *aumentar o grau de semelhança* entre os casos e os controles. Desse modo, ele será responsável por identificar registros na base de controle com características semelhantes aos registros da base de análise (os casos), como por exemplo etnia, sexo, município, tipo de exame, período, entre outras características.

1.3 Trabalhos Relacionados

Conforme descrito anteriormente, a proposta deste trabalho implica em definir um *grupo de análise* derivado da base SIVEP-Gripe e um *grupo de controle* derivado da base e-SUS. Para isso, propõe-se a elaboração de um algoritmo de *pareamento* responsável por encontrar para cada registro do SIVEP-Gripe outros registros do e-SUS que compartilhem

algumas características como idade, sexo, etnia, entre outras características. O objetivo final é conseguir verificar a *prevalência de genes alelos* na população que teve COVID-19.

Neste trabalho é realizado um ECC (RÊGO, 2010), técnica que vem sendo cada vez mais utilizada nas áreas da Epidemiologia e da Saúde Pública. Em resumo, o ECC seleciona indivíduos que apresentam ou não uma determinada condição (nesse caso uma doença). Essa é uma abordagem diferente da adotada pelo estudo de *coorte* que inicialmente seleciona os indivíduos com base em uma exposição a um fator de risco para uma doença.

Segundo (RÊGO, 2010), os casos podem ser selecionados ou da população hospitalar ou da população em geral com auxílio de serviços de vigilância epidemiológica. A seleção do grupo de controle pode ser feita pela técnica do pareamento que toma como base *variáveis* que sejam *fatores de risco* para a doença estudada (RÊGO, 2010). Desse modo, para cada caso são selecionados um ou mais controles conforme normas pré-estabelecidas.

Alguns trabalhos realizados na literatura já realizaram algum tipo de pareamento não necessariamente sendo por um algoritmo, mas podendo ser realizado manualmente pelos pesquisadores. (PEREIRA et al., 2016) propõe um ECC na qual realiza-se um pareamento por idade e sexo para inferir a relação entre diabetes e tuberculose. Por sua vez, Lopes e Coutinho (LOPES; COUTINHO, 1999) avaliaram o quanto os transtornos mentais e a dependência ao álcool são fatores de risco para a dependência de cocaína utilizando também um pareamento por idade e sexo além de aplicar uma regressão logística condicional, que reteve o pareamento além de permitir o ajuste com relação a variáveis não utilizados no pareamento. Neste estudo, considerou-se também algumas variáveis sociodemográficas como por exemplo: nº de anos na escola, situação marital, se é pai ou mãe e situação ocupacional. Por último, vale mencionar o estudo (MACHADO et al., 2022), que elabora um ECC para entender o impacto de infecções contraídas em prol da assistência à saúde no Brasil. Nele é proposto um pareamento manual, selecionando um controle para cada caso, no qual o grupo de controle é formado pelos indivíduos que ainda não haviam contraído infecção. Neste ECC, o pareamento considera como variáveis, além da idade e do sexo, o motivo da internação e o tempo de risco (permanência total no hospital).

Por último, é importante ressaltar que não foram encontrados na literatura trabalhos que proponham um ECC com um pareamento baseado em algum algoritmo que seja semelhante ao exposto por este trabalho. As referências encontradas apresentam um pareamento feito de maneira manual pelos pesquisadores, sendo que alguns deles também utilizam a idade e o sexo como variáveis.

2 METODOLOGIA

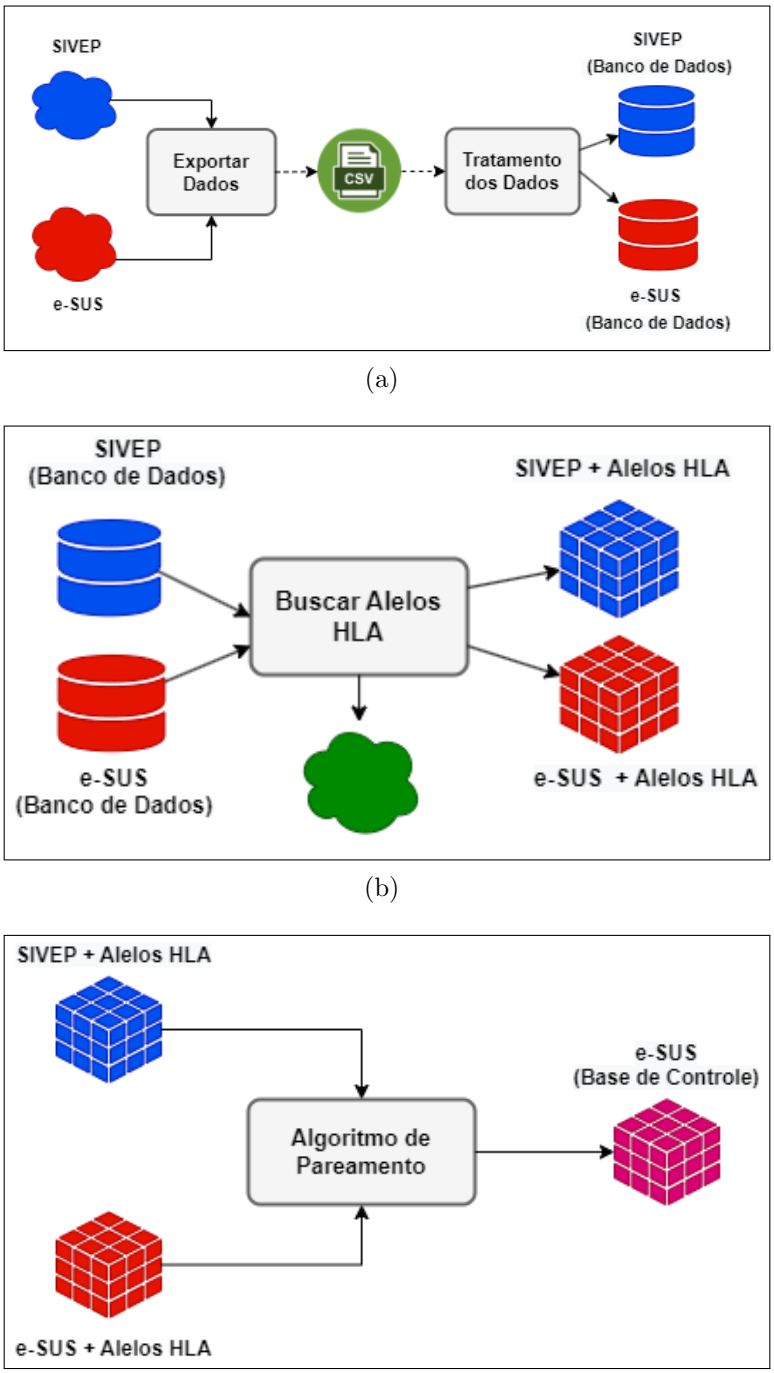
O estudo foi aprovado pela CONEP (Comissão Nacional de Ética em Pesquisa) (CAAE - 40921320.1.0000.5259). Os pesquisadores envolvidos assinaram termo de confidencialidade sobre o uso dos dados pessoais utilizados para cruzamento das bases de dados e a base de estudo gerada é oferecida para uso já com uma *codificação anonimizada dos registros*, em consonância também com a LGPD (Lei Geral de Proteção de Dados).

A metodologia utilizada para criação de uma base de ECC, para a análise da prevalência de alelos em pacientes com COVID-19, pode ser dividida em três etapas, que podem ser visualizadas na Figura 6:

A etapa 1, Figura 6a, é crucial para este trabalho, uma vez que ela define as bases de dados de onde serão extraídos os pacientes que tiveram casos confirmados ou suspeitos de COVID-19. Conforme explicado na Seção 1.1, dois SIS foram de suma importância durante a pandemia de COVID-19 para ajudar a monitorar os casos da doença. O SIVEP-Gripe, que é utilizado para armazenar todas as notificações referentes a SRAG, encerradas ou não. O e-SUS que foi criado durante a pandemia para registrar casos suspeitos ou confirmados especificamente de COVID-19. Esta primeira etapa intitulada de *pré-processamento* se caracteriza, inicialmente, pela exportação dos dados armazenados nas bases externas do sistemas SIVEP-Gripe e e-SUS para arquivos (por exemplo, no formato *.csv*). A exportação dos dados para arquivos é uma operação feita por *usuário autenticado* com acesso aos dois sistemas. Em seguida, os dados presentes nestes arquivos passaram por um *tratamento* no qual envolveu uma *uniformização* e *limpeza* para serem então *persistidos e integrados a um banco de dados relacional*. Os dados persistidos foram mantidos *anonimizados* levando-se em consideração a LGPD. Formam-se assim as duas bases locais com os dados do SIVEP-Gripe e e-SUS usados nas etapas posteriores. As bases contém apenas informações sobre a situação dos pacientes com sintomas ou que tiveram a COVID-19, como por exemplo, o resultado do teste, o tipo do teste, a data de notificação, entre outras.

A etapa 2, Figura 6b, apresenta a abordagem para se conseguir os alelos HLA desses pacientes. Uma possibilidade seria realizar o exame de tipificação HLA de todos (ou um grupo) de pacientes das bases SIVEP-Gripe e e-SUS. Contudo, essa abordagem é impraticável por dois motivos principais. Primeiro, o alto custo envolvido na realização de exame de tipificação HLA (i.e. exame que identifica os alelos de um paciente) para cada um dos pacientes que serão utilizados na pesquisa. A segunda razão que torna essa abordagem inviável é que muitos pacientes presentes nessas bases faleceram de COVID-19. Assim, a abordagem proposta para se conseguir os alelos HLA para os pacientes das bases SIVEP-Gripe e e-SUS foi buscar esses dados no REDOME, Figura 6b, uma vez que os alelos HLA dos doadores é a principal informação armazenada nesta base de dados.

Figura 6 - Etapas da Metodologia Adotada.



Legenda: (a) Pré-processamento dos Dados. (b) Aquisição dos Alelos HLA. (c) Criação da Base de Controle.

Assim os alelos dos pacientes das bases SIVEP-Gripe e e-SUS foram obtidos no banco de dados do REDOME.

A última etapa, Figura 6c, consiste em gerar a base de controle, em função das características dos pacientes da base de casos. A abordagem proposta é a criação de um algoritmo de pareamento que escolhe os registros para a base de controle de acordo com as características dos registros da base de casos. Neste trabalho, como a base SIVEP-Gripe é composta de casos graves incluindo aqueles que vieram a óbito, ela foi escolhida para ser a base de casos, enquanto que a base e-SUS, composta de casos suspeitos e confirmados de COVID-19, foi escolhida para ser a base de controle. A utilização de um algoritmo para escolha dos registros tem dois objetivos. Em primeiro lugar, criar uma base de controle com características similares à base de casos. Em segundo lugar, evitar que o pesquisador tenha que participar da escolha dos registros de forma não randômica no pareamento entre casos e controles.

A seguir são descritas as implementações de cada uma das etapas apresentadas.

2.1 **Pré-Processamento dos Dados**

Como mostrado na Seção 1.1, SIVEP-Gripe e e-SUS são sistemas de informação criados/incorporados para notificar casos de COVID-19. Ambos os sistemas possuem uma página web no qual o acesso depende de uma autenticação do usuário fornecendo informações como nome, email ou senha. Vale mencionar novamente que:

- casos graves e internados de SRAG, causados pela COVID-19, são notificados no SIVEP-Gripe;
- casos suspeitos ou confirmados de SG são notificados no e-SUS, além de se registrar nele também o tipo de teste realizado para detectar a doença (RT-PCR ou sorológico);

É possível se cadastrar nestes sistemas sendo necessário que o usuário informe alguns de seus dados pessoais. Tendo cadastro concluído, o usuário poderá acessar os registros de notificações reportadas nestes sistemas que são alimentadas por profissionais e unidades de saúde. Desse modo, a *consistência* das informações fica dependendo do preenchimento correto dos dados.

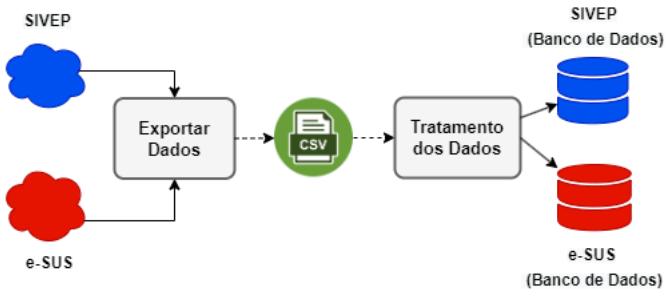
Pensando-se no cenário de controle epidemiológico da COVID-19 faz sentido querer manter um nível de consistência entre as notificações cadastradas nestes sistemas. Profissionais e instituições que estudam e pesquisam sobre o vírus necessitam que as fontes de dados sejam confiáveis, possuindo o maior grau de assertividade possível. Assim, é importante que os dados possam passar por um processo de *uniformização e limpeza que*

os padronize e que possa *facilitar na integração das informações*. Visa-se facilitar o manuseio dos dados, uma vez que estes são preenchidos em hospitais, UPAs (Unidades de Pronto Atendimento) e postos de saúde, onde, muitas vezes, as informações guardadas não são conferidas através de documentos, estando em situações de emergência e, desse modo, muito sujeita a erros de digitação.

Como dito anteriormente, os dados foram exportados das bases externas dos sistemas SIVEP-Gripe e e-SUS para serem então persistidos em duas bases locais num banco de dados relacional. Devido a algumas *anomalias* ou *inconsistências* encontradas nos dados, elaborou-se uma etapa de *pré-processamento* que implicou num *tratamento* responsável pela eliminação ou redução destes problemas. Esse tratamento dos dados envolveu uma uniformização e limpeza para poderem ser então integrados e persistidos no banco de dados. Assim, foi feita uma análise para investigar possíveis inconsistências entre os dados sendo que as mais comuns encontradas foram aquelas relacionadas a *erros de digitação*, *duplicidade de registros* (dois ou mais registros de notificações referentes ao mesmo paciente) e alguns outros tipos de divergências encontradas entre os registros de notificações referentes ao mesmo paciente (existiam, por exemplo, dois ou mais registros de notificações do mesmo paciente apresentando endereços distintos).

Essa fase de pré-processamento dos dados responsável pela uniformização e limpeza pode ser visualizada na Figura 7:

Figura 7 - Pré-Processamento dos Dados.



Como ilustrado pela Figura 7, os seguintes passos foram realizados:

- primeiramente, os dados dos sistemas SIVEP-Gripe e e-SUS foram salvos e disponibilizados em arquivos *.csv*. Para isso, um usuário cadastrado em ambos os sistemas se autenticou neles, selecionou as notificações e exportou seus dados em formato *.csv*;

- em seguida, esses dados foram importados para uma aplicação responsável por uniformizar estes dados e gerar uma padronização de formatação entre eles. Dessa forma, foi possível realizar outros ajustes para limpeza dos dados identificando e removendo tanto as duplicidades como as inconsistências encontradas entre eles;

- ao final, os dados foram persistidos em um banco de dados relacional.

Detalhes sobre os principais tipos de ajustes aplicados sobre os dados provenientes das bases SIVEP-Gripe e e-SUS são apresentados no Apêndice 3.3.

2.2 Aquisição dos Alelos HLA

O objetivo do pré-processamento apresentado na seção anterior foi de construir bases locais consistentes a partir dos dados encontrados nos registros de notificações retirados dos sistemas SIVEP-Gripe e e-SUS e, com isso, poderem ser utilizadas com segurança para estudar características da COVID-19. Assim, após todo este processo realizado o pesquisador pode confiar nos dados disponíveis. Os dados consolidados e produzidos pelo pré-processamento foram persistidos num banco de dados relacional além de serem anonimizados conforme a LGPD.

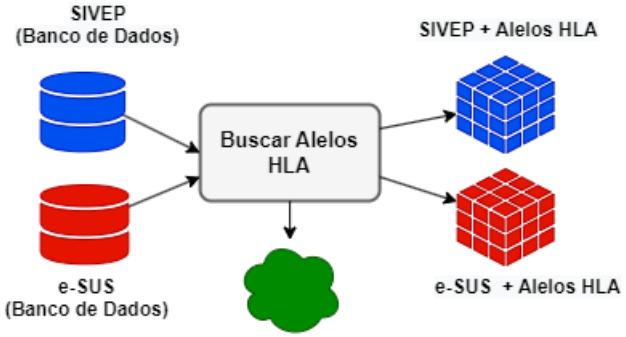
O objetivo final deste trabalho é identificar a *prevalência de alelos HLA* em pacientes com a COVID-19. Porém, as bases SIVEP-Gripe e e-SUS contêm apenas informações gerais sobre pacientes com sintomas ou que tiveram a COVID-19. Não há nenhuma informação sobre os alelos HLA, necessários para realização desta avaliação. Assim, ao invés de ter que realizar a tipagem HLA destes pacientes para se conseguir as informações sobre os alelos, o que seria muito custoso (o custo do exame de tipificação HLA de cada registro das bases SIVEP-Gripe e e-SUS e, em muitos casos, inviável (muitos pacientes faleceram), as informações sobre os alelos foram buscadas na base de dados do REDOME (Brasil, Ministério da Sáude, f), que contêm mais de 5 milhões de voluntários. A principal informação armazenada no REDOME é a tipificação HLA dos doadores que é utilizada para verificar a compatibilidade entre paciente e doador.

O campo raça/etnia tem uma importância muito grande na avaliação de alelos HLA. Assim, como a informação original sobre a raça/etnia das bases SIVEP-Gripe e e-SUS estava bastante incompleta, essa informação foi buscada na base do REDOME. A etnia/raça foi um dos campos mais observados para analisar a prevalência de alelos em pacientes que tiveram contato com a COVID-19, pois ela tem uma influência direta no alelo que pode ser encontrado (NUNES et al., 2020). Por isso era essencial que as informações sobre etnia/raça fossem as mais consistentes possíveis para poderem ser então utilizadas pelo algoritmo de pareamento.

Conforme pode ser visto na Figura 8, a abordagem proposta para se conseguir os alelos HLA para os pacientes das bases SIVEP-Gripe e e-SUS foi buscar estes dados no REDOME. Esta informação está presente no REDOME, pois a compatibilidade HLA entre paciente-doador é o principal fator para o sucesso do transplante de medula óssea (Silva, Marcio NP and Cristóvão, Luís and Pôrto, MS and Marzulo, Leandro AJ and Sena, Alexandre C, 2019). Ao final, foram geradas novas bases locais SIVEP-Gripe e

e-SUS com todas as informações anteriores já persistidas no banco de dados e também com os dados sobre os alelos dos pacientes. A base de dados do REDOME disponibilizada para consulta, só permitia a validação das seguintes variáveis para associação: nome, data de nascimento, raça/etnia, nome da mãe, UF e código IBGE do município de moradia na data do cadastro.

Figura 8 - Aquisição dos Alelos HLA.



2.3 Criação da Base de Controle

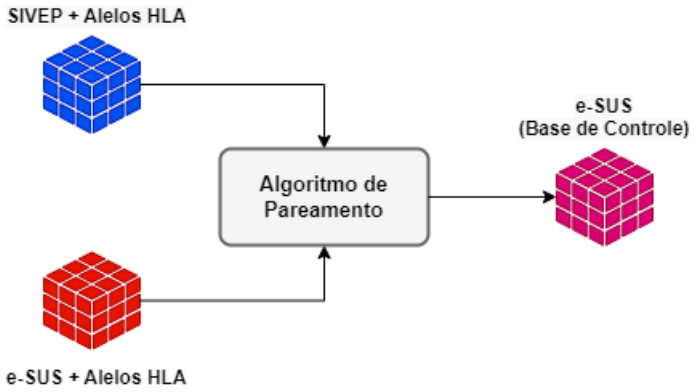
Uma vez que o objetivo final da base de dados a ser gerada é analisar a prevalência de alelos HLA em pacientes que tiveram COVID-19, é fundamental para o estudo ter uma base de casos a ser analisada e uma base de controle. Neste trabalho, a base a ser analisada tem a origem dos dados no SIVEP-Gripe, uma vez que os pacientes dessa base são compostos por casos graves de infecções ou óbitos causados pelo SARS-CoV-2. Por sua vez, a base de controle será composta por dados provenientes do e-SUS, que contêm casos suspeitos ou confirmados de COVID-19.

Para se conseguir uma base de dados final consistente para qualquer estudo clínico, é necessário que ela seja *homogênea*. Ou seja, para cada registro da base de análise se consiga 2 ou mais registros com características semelhantes de uma outra base para servir como controle. Assim, este trabalho propõe e implementa um algoritmo para realizar o pareamento entre uma base de análise com uma base de controle. O pareamento consiste em encontrar para cada registro do SIVEP-Gripe um ou mais registros do e-SUS que apresentem semelhanças considerando características como: etnia, sexo, localização, resultado de teste, tipo de teste realizado, entre outras.

Conforme ilustrado pela Figura 9, a base de controle (e-SUS) é gerada em função das características dos pacientes escolhidos para a base de casos (SIVEP-Gripe). Os registros do SIVEP-Gripe são pareados com alguns registros do e-SUS. Os registros do e-SUS pareados e selecionados pelo algoritmo são então armazenados na base de controle.

É esperado que os dados gerados pelo algoritmo de pareamento proposto neste trabalho tornem a base final a mais homogênea possível, minimizando a interferência das diferentes características entre as bases de dados nos estudos clínicos/estatísticos a serem realizados.

Figura 9 - Criação da Base de Controle.



2.3.1 Objetivo

O objetivo do algoritmo é parear cada registro do SIVEP-Gripe com uma quantidade $2 \times N$ de registros do e-SUS sendo: N registros com resultado do teste de COVID-19 positivo e N registros com resultado do teste de COVID-19 negativo. Para isso, é importante parear registros que tenham características similares pois, como dito anteriormente, pretende-se selecionar registros do e-SUS que possam servir como grupo de controle e de testes experimentais para, por exemplo, um estudo de caso como a identificação da ocorrência de alelos em casos confirmados de COVID-19.

Os campos escolhidos para serem pareados, em função das características das bases de dados do SIVEP-Gripe e do e-SUS, foram:

- *idade*: usado para parear os registros por faixa etária já que seria difícil obter a quantidade mínima de registros pareados que tivessem idade exata;
- $resultadoTeste$: indica se o paciente teve ou não COVID-19;
- $tipoTeste$: indica o tipo de teste realizado pelo paciente para saber se teve ou não COVID-19;
- $dataNotificacao$: indica a data em que o paciente foi notificado em alguma das bases e é a informação usada para parear os pacientes por período estabelecido pelo algoritmo;
- *sexo*: usado para indicar o sexo do paciente;

- *etniaRedome*: campo capturado do REDOME que indica a raça/etnia do paciente;
- *municipio/regiao*: usados para agrupar e parear os registros por localidade.

Uma vez que a distribuição dos alelos varia bastante dependendo da etnia do paciente, ela é a característica mais relevante para uma análise de alelos HLA. A data de notificação determina o dia em que o paciente procurou a unidade de saúde. É importante que, ao escolher um paciente do grupo de controle (e-SUS), a data de notificação seja próxima da data de notificação do paciente do grupo de casos (SIVEP-Gripe) para, por exemplo, aumentar a chance deles terem sido infectados pela mesma variante da COVID-19. A localidade que o paciente reside também é importante. Assim, o ideal é escolher pacientes que residam em um mesmo município ou, pelo menos, na mesma região. O município/região é usado para tentar escolher registros da base de controle possivelmente infectados pela mesma variante da COVID-19 e também porque municípios pequenos podem ter uma distribuição alélica diferente da distribuição esperada. Por exemplo, dependendo do município pode haver uma quantidade variável para cada tipo de raça/etnia. Também é importante que a distribuição do sexo da base de controle seja similar à base de casos. Por fim, o ideal seria que todos os pacientes tivessem sido testados através de PCR. Porém, como não é possível, o algoritmo prioriza o RT-PCR, depois o Teste Rápido Antígeno e, por fim, o Teste Rápido Anticorpo.

Outra característica muito importante é a idade do paciente. Porém, encontrar pacientes na base e-SUS com exatamente a mesma idade não seria viável. Logo, os registros da base de casos SIVEP-Gripe foram divididos em 3 faixas etárias distintas através do cálculo dos percentis 33 e 67. Ou seja, a primeira faixa etária é composta do valor da menor idade da base até o valor do percentil 33% (o valor do percentil é arredondado para baixo). A segunda faixa etária é composta do valor posterior ao percentil 33% (arredondado para cima) até o valor do percentil 67% (arredondado para baixo). Por fim, a terceira faixa etária é composta do valor posterior ao percentil 67% (arredondado para cima) até o valor da maior idade da base. No momento do pareamento apenas as idades dentro das faixas etárias são selecionadas.

Inicialmente o algoritmo baseia-se apenas nestes campos mas, ele foi elaborado de tal forma que possa ser extensível para outros campos que queira-se utilizar no pareamento. Deve-se mencionar também que o algoritmo agrupa registros do SIVEP-Gripe fazendo o pareamento com registros do e-SUS de forma isolada para cada um dos possíveis valores do campo *evolucaoCaso*: “OBITO”, “UTI”, “INTERNADO” e “RECUPERADO”.

2.3.2 Algoritmo de Pareamento

Nesta subseção, será explicada com mais detalhes a lógica utilizada para construir o algoritmo de pareamento. Serão apresentadas as filtragens aplicadas sobre os registros do e-SUS com relação a cada registro do SIVEP-Gripe. Conforme mostrado pelo Algoritmo 1, são passados como parâmetros de entrada os arquivos com os registros do SIVEP-Gripe ($arqSivep$) e do e-SUS ($arqSus$). Ao final, é produzido um arquivo final de pareamento ($arqPar$) contendo todos os registros do SIVEP-Gripe juntamente com os registros do e-SUS pareados pelo algoritmo.

Algoritmo 1 - Algoritmo de Pareamento.

DOCUMENTAÇÃO

TÍTULO

Algoritmo de Pareamento

PROPÓSITO

Parear registros do SIVEP-Gripe com o e-SUS.

ENTRADAS

arqSivep: lista de registros do SIVEP-Gripe
arqSus: lista de registros do e-SUS
N: tamanho do pareamento

SAÍDAS

listaSus: lista de registros pareados do e-SUS

ALGORITMO PAREAMENTO

1. **declarar** $arqSivep, arqSus, listaSus, listaPos, listaNeg$ **listas**
2. **declarar** reg **registro**
3. **declarar** $N, filtro, semanas$ **numéricos**
4. $filtro \leftarrow 1$
5. $semanas \leftarrow 1$
6. $listaPos \leftarrow \emptyset$
7. $listaNeg \leftarrow \emptyset$
8. $listaPar \leftarrow \emptyset$
9. **para** reg **de** $arqSivep[1]$ **até** $arqSivep[N]$, **fazer**
10. **enquanto** $((listaPos.tam < N$ *ou* $listaNeg.tam < N)$ *e* $filtro < 10)$, **fazer**
11. $listaSus \leftarrow arqSus - listaPar$
12. **se** $(filtro < 9)$, **então**
13. $listaSus \leftarrow$ FILTRAR$(listaSus, reg.etniaRedome)$
14. **fim se**
15. **se** $(filtro < 8)$, **então**
16. $listaSus \leftarrow$ FILTRAR$(listaSus, reg.dataNotificacao, semanas)$
17. **fim se**

— *continua* —

Algoritmo 1 - Algoritmo de Pareamento. (continuação)

— *continuação* —
se $(filtro < 4)$, **então**
$listaSus \leftarrow$ FILTRAR$(listaSus, reg.municipio)$
fim se
se $(filtro == 4)$, **então**
$listaSus \leftarrow$ FILTRAR$(listaSus, reg.regiao)$
fim se
se $(filtro < 3)$, **então**
$listaSus \leftarrow$ FILTRAR$(listaSus, reg.sexo)$
fim se
se $(filtro < 2)$, **então**
$listaSus \leftarrow$ FILTRAR$(listaSus, reg.tipoTeste)$
fim se
se $(listaPos.tam < N)$, **então**
$listaPos \leftarrow listaPos \cup$ RECEBEPOSITIVOS$(listaSus, N)$
fim se
se $(listaNeg.tam < N)$, **então**
$listaNeg \leftarrow listaNeg \cup$ RECEBENEGATIVOS$(listaSus, N)$
fim se
se $(filtro > 4\ e\ filtro < 8)$, **então**
$semanas \leftarrow semanas + 1$
fim se
se $(listaPos.tam == N\ e\ listaNeg.tam == N)$, **então**
$filtro \leftarrow 10$
fim se
$filtro \leftarrow filtro + 1$
fim enquanto
$listaPar \leftarrow listaPar\ \cup\ listaPos$
$listaPar \leftarrow listaPar\ \cup\ listaNeg$
fim para
FIM ALGORITMO
FIM DOCUMENTAÇÃO

O pareamento é realizado das linhas 1 à 40 do algoritmo para cada registro, *reg*, da base SIVEP-Gripe. Na linha 2, as variáveis $filtro$ e *semanas* são inicializadas. As listas que guardam os registros positivos encontrados ($listaPos$), os registros negativos encontrados ($listaNeg$) e a lista final contendo todos os registros da base de controle a ser gerada ($listaPar$) são inicializadas com valor nulo (linha 3).

O processo de pareamento de um registro acontece das linhas 4 à 37 através de um laço do tipo *enquanto*. Esta ação só termina quando forem encontrados N registros positivos e N registros negativos ou se não existirem N registros mesmo após todos os filtros serem liberados. Para cada registro a ser pareado, a lista de registros do e-SUS que podem ser selecionados ($listaSus$) é inicializada com todos os registros da base e-SUS

($arqSus$) menos os registros que já foram selecionados pelo algoritmo de pareamento que ficam armazenados na lista $listaPar$ (linha 5). Ou seja, inicialmente todos os registros serão selecionados pois a lista $listaPar$ está vazia.

A próxima etapa do algoritmo é aplicar os filtros que definem as características a serem pareadas. A aplicação ou não do filtro é definida pela variável $filtro$, inicializada com valor 1. Ou seja, inicialmente todos os filtros serão aplicados. Assim, a variável que contém os registros do e-SUS que podem ser selecionados ($listaSus$) primeiro é filtrada pelo campo *etniaRedome* (linhas 6 à 8) presente no registro SIVEP-Gripe (e.g. apenas registros com campo *etniaRedome* "PARDA" serão selecionados caso o campo *etniaRedome* do registro SIVEP-Gripe seja "PARDA"). Em seguida, é aplicado o filtro pela data de notificação (linhas 9 à 11). Ou seja, inicialmente, apenas os registros que tiverem data de notificação uma semana para frente ou para trás da data de notificação do registro do SIVEP-Gripe serão selecionados. Repare que esse filtro é aplicado em cima do filtro do campo *etniaRedome* (e.g. apenas registros da etnia "PARDA" no período de 1 semana para frente ou para trás da data de notificação serão selecionados). O próximo filtro a ser aplicado é o do campo *municipio* que seleciona apenas registros do mesmo município que o registro do SIVEP-Gripe (linhas 12 à 14). Como selecionar pacientes de um mesmo município pode ser inviável, o algoritmo está preparado para, caso não encontre os N registros positivos e negativos de um mesmo município, filtrar por região (linhas 15 à 17). O processo de filtragem continua aplicando o filtro sobre o campo *sexo* (linhas 18 à 20) e, depois, sobre o campo $tipoTeste$ (linhas 21 à 23).

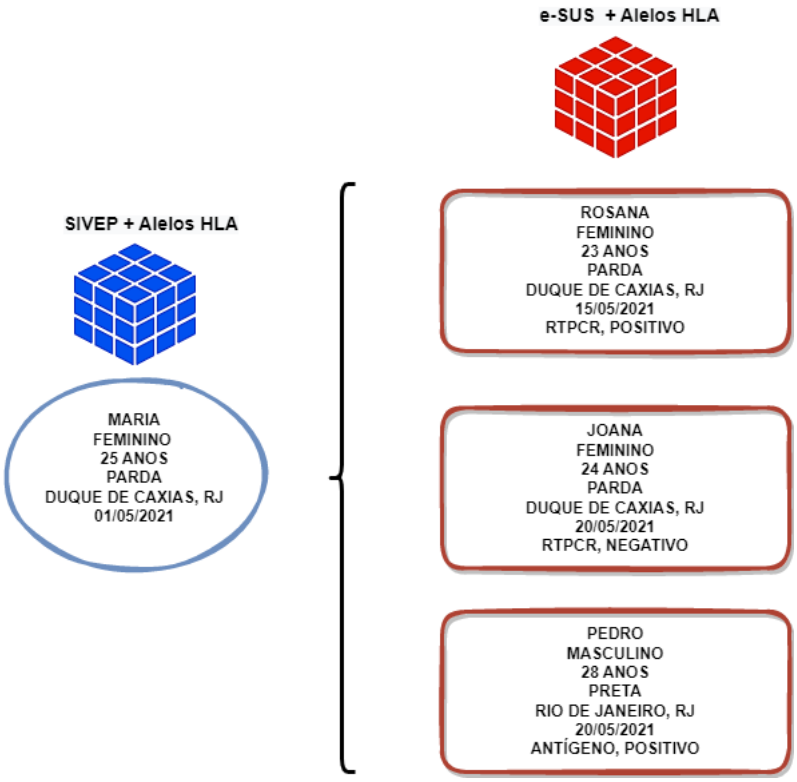
Em seguida, o algoritmo armazena na lista de positivos ($listaPos$) os N primeiros registros positivos após toda a filtragem (linhas 24 à 26) e na $listaNeg$ os primeiros registros negativos após toda a filtragem. Caso tenha encontrado registros suficientes, o pareamento do registro atual termina atribuindo o valor 10 para variável $filtro$ (linhas 30 à 32). Repare que neste caso o pareamento realizado foi extremamente adequado, pois todos os filtros desejados foram aplicados. Caso contrário, o pareamento do registro atual irá continuar, porém, com um filtro a menos pois a variável $filtro$ é incrementada de 1 (linha 36). Nesse caso, o processo irá continuar até conseguir a quantidade de registros positivos e negativos desejada, sendo que a cada nova iteração um novo filtro é retirado. A partir do valor 5 para a variável $filtro$ a semana de notificação é incrementada (linhas 33 à 35). Ou seja, o período para selecionar a data de notificação é aumentado para 2 semanas para frente ou para trás, podendo chegar até 4 semanas. É importante destacar que quantos mais filtros forem retirados mais desbalanceado será o pareamento.

2.3.2.1 Exemplo de Pareamento

Para demonstrar como é feito um pareamento pelo Algoritmo 1 apresentado na Subseção 2.3.2, considere o exemplo com dados de pacientes fictícios mostrados na Figura 10 que ilustra um pareamento entre um registro do SIVEP-Gripe com outros três registros do e-SUS. Em um pareamento perfeito, para cada registro do SIVEP-Gripe seriam selecionados do e-SUS registros que tivessem total compatibilidade entre as características adotadas pelo algoritmo. Neste exemplo, tem-se um registro do SIVEP-Gripe no qual:

- o nome do paciente é "MARIA";
- o sexo é "FEMININO";
- a idade é "25 ANOS";
- a raça ou etnia é "PARDA";
- o município de notificação é "DUQUE DE CAXIAS";
- a data de notificação é "01/05/2021".

Figura 10 - Exemplo de pareamento entre registros do SIVEP-Gripe e e-SUS.



A meta é encontrar registros do e-SUS que tenham preferencialmente os mesmos valores para cada um destes campos. Assim, para este registro do SIVEP-Gripe o ideal seria encontrar registros do e-SUS que tenham:

- a idade na mesma faixa etária estabelecida pelos cálculos dos percentis 33 e 67;
- a raça ou etnia "PARDA";
- notificado no município de "DUQUE DE CAXIAS" ou em outro município pertencente à mesma região;
- a data de notificação compreendida em até uma semana para frente ou para trás da data "01/05/2021".

No exemplo ilustrado pela Figura 10, os dois primeiros registros selecionados (com os nomes "ROSANA" e "JOANA") do e-SUS tem compatibilidade de valores com todos os correspondentes no SIVEP-Gripe. Contudo, para o 3º registro do e-SUS (com o nome "PEDRO") vale mencionar algumas observações:

- o valor "MASCULINO" do sexo não corresponde ao valor "FEMININO" do sexo no SIVEP-Gripe;
- o valor "PRETA" da raça/etnia não corresponde ao valor "PARDA" da raça/etnia no SIVEP-Gripe;
- o valor "RIO DE JANEIRO, RJ" do município não corresponde ao valor "DUQUE DE CAXIAS, RJ" do município no SIVEP-Gripe. Contudo, apesar de haver uma divergência nos valores do município, o algoritmo conseguiu parear um registro do e-SUS cujo município encontra-se na mesma região ao daquele encontrado no SIVEP-Gripe;
- foi pareado pelo algoritmo mesmo apresentando o valor "ANTÍGENO" no campo tipo de teste realizado apesar de a preferência ser em selecionar registros com o valor "RTPCR" para este campo;
- sendo a data de notificação "20/05/2021" o algoritmo também conseguiu parear este registro por este campo visto que é próxima o suficiente da data de notificação "01/05/2021" do registro do SIVEP-Gripe mantendo-se na margem de um mês para frente ou para trás.

3 RESULTADOS E DISCUSSÃO

Neste capítulo são descritos os resultados da avaliação do algoritmo proposto como também da análise da prevalência de alelos HLA em pacientes com COVID-19 das bases SIVEP-Gripe e e-SUS. Os resultados foram avaliados com a utilização do *software* estatístico Epi Info projetado pelo CDC (Centers for Disease Control and Prevention) para a comunidade global de médicos e pesquisadores da saúde pública (Epi Info,).

Para poder comparar os resultados produzidos pelo algoritmo de pareamento proposto, foram criados outros dois algoritmos. O algoritmo chamado de **Sem Filtro** é basicamente o mesmo algoritmo apresentado, porém sem aplicar nenhum filtro. Ou seja, ele escolhe os N primeiros registros positivos e negativos disponíveis para cada registro do SIVEP-Gripe sem aplicar nenhum filtro. Por outro lado, o algoritmo chamado de **Aleatório** seleciona para cada registro do SIVEP-Gripe N registros positivos e negativos aleatoriamente. Ainda, para verificar o desempenho do algoritmo de acordo com o tamanho de N, cada um dos três algoritmos foi executado com os valores 2, 3 e 4.

Utilizando as bases SIVEP-Gripe e e-SUS geradas na etapa de aquisição dos alelos (ver Seção 2.2), cada algoritmo produz uma base de controle contendo apenas os registros do e-SUS selecionados pelo algoritmo de pareamento a partir dos registros encontrados na base de casos SIVEP-Gripe. Ao final, é produzido um único arquivo *.csv* com todos estes registros sendo que a quantidade varia conforme o tamanho de N passado como entrada ao algoritmo. Por fim, é importante observar que a escolha do algoritmo de pareamento fica limitada à base e-SUS com os dados dos alelos que contém os registros para a criação da base de controle. Ou seja, de acordo com os registros disponíveis, o arquivo *.csv* final produzido contém o melhor pareamento que o algoritmo consegue produzir. Para a identificação dos registros, também foi acrescentado ao arquivo um campo chamado *origem* que informa se o registro é proveniente do SIVEP-Gripe ou do e-SUS. Nas seções a seguir, serão apresentados os resultados separadamente para cada uma das duas bases avaliadas, referentes aos estados do Rio de Janeiro e Minas Gerais. Mais especificamente, as seções 3.1 e 3.2 avaliam o algoritmo de pareamento para as bases do RJ e MG, respectivamente, enquanto a seção 3.3 avalia a prevalência de alelos HLA para pacientes com COVID-19 do estado de Minas Gerais. É importante ressaltar que ainda não foi realizada a análise da prevalência para a base do Rio de Janeiro pois, recentemente, foram recebidos novos arquivos com uma quantidade muito maior de dados que permitirá uma análise mais precisa, uma vez que a base original do RJ tem apenas 577 registros na base de casos, o que é considerado insuficiente.

3.1 Avaliação do Algoritmo de Pareamento RJ

Esta seção avalia o algoritmo de pareamento utilizando às bases SIVEP-Gripe e e-SUS do estado do Rio de Janeiro. Após todo o pré-processamento e busca dos alelos HLA no REDOME essas bases ficaram compostas com as seguintes quantidades de registros:

- o SIVEP-Gripe continha 577 registros;
- o e-SUS continha 26.546 registros.

A quantidade de registros do arquivo final correspondente a base de controle após o pareamento depende do valor do N escolhido. Por exemplo, para $N = 3$ o arquivo de casos terá $2 \times 3 \times 577 = 3.462$ registros, onde 1.731 registros são positivos para COVID-19 e 1.731 são negativos.

As subseções a seguir apresentam a avaliação da homogeneidade para cada um dos campos selecionados. Além disso, é importante destacar novamente que a avaliação foi realizada para 3 valores de N distintos: 2, 3 e 4. Também é válido ressaltar que o valor de N deve ser escolhido pelo usuário, de acordo com a quantidade de registros do seu ECC.

3.1.1 Avaliação do Campo Idade

Inicialmente, foi avaliada a distribuição do campo *idade*. Conforme descrito na Subseção 2.3.2, a base SIVEP-Gripe foi dividida em 3 faixas etárias, de maneira que de acordo com a *idade* do registro do SIVEP-Gripe apenas os registros do e-SUS com o campo *idade* dentro da faixa etária prevista ficassem disponíveis. A menor idade, maior idade e idade média para cada uma das bases de controle geradas pelos algoritmos podem ser vistas na Tabela 1. Além disso, a tabela apresenta a idade obtida para cada valor de N (e.g. $N = 2$ é representado por $2:1$).

Tabela 1 - Valores máximo, médio e mínimo para o campo *idade*

Idade	**SIVEP**	**e-SUS**								
		Com Filtro			**Sem Filtro**			**Aleatório**		
		2:1	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**
Mínima	22	23	22	22	19	19	19	20	19	19
Média	49	48	47	48	40	40	41	42	41	41
Máxima	75	73	73	73	72	72	72	69	74	74

Quando se compara os valores gerados por cada um dos três algoritmos com os valores do SIVEP-Gripe, fica claro o melhor desempenho do algoritmo proposto (**Com Filtro**). Não só a idade mínima e máxima ficaram muito próximas, mas, principalmente a

média aritmética das idades. Com relação a variação do valor de N, não houve mudança significativa. Ao se comparar os valores para os outros dois algoritmos, o algoritmo **Aleatório** produziu resultados ligeiramente melhores que o algoritmo **Sem Filtro**.

3.1.2 Avaliação do Campo Etnia

O campo mais importante do algoritmo de pareamento proposto é o referente à raça/etnia do paciente, uma vez que essa característica tem uma influência significativa no alelo que pode ser encontrado (NUNES et al., 2020). Em função disso, o primeiro filtro a ser aplicado é pelo campo *etniaRedome*, conforme descrito na Subseção 2.3.2. Ou seja, ele é o último filtro a ser ignorado no algoritmo de pareamento. Os resultados da frequência (em porcentagem) para cada uma das etnias presentes nas bases de controle geradas pelos algoritmos, assim como o valor do chi-quadrado, podem ser visualizados na Tabela 2:

Tabela 2 - Valores em percentual (%) para frequência dos valores para o campo *etniaRedome*

Etnia	**SIVEP**	**e-SUS**								
		Com Filtro			**Sem Filtro**			**Aleatório**		
		2:1	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**
Amarela	1.39	1.43	1.42	1.43	1.65	1.39	1.34	1.30	1.30	1.47
Branca	59.62	59.88	60.51	60.66	64.25	63.78	63.58	62.61	62.71	61.63
Ignorado	0.0	0.0	0.0	0.11	1.47	1.42	1.67	1.56	1.73	1.71
Indígena	1.04	1.08	0.92	0.76	0.35	0.35	0.41	0.43	0.23	0.37
Não Inf.	2.95	2.56	1.99	1.58	0.17	0.35	0.45	0.26	0.35	0.19
Parda	18.89	18.93	19.04	19.24	18.80	19.61	19.95	21.36	20.85	20.93
Preta	16.12	16.12	16.12	16.23	13.30	13.11	12.59	12.48	12.82	13.69
chi-2		0.28	2.25	6.78	66.23	64.83	63.83	60.51	71.63	96.19
p		99.8%	81.3%	34.1%	0%	0%	0%	0%	0%	0%

Ao se comparar os valores de frequência para cada uma das etnias das bases de controle (e-SUS) geradas pelos algoritmos de pareamento com os valores da base de casos (SIVEP-Gripe) pode-se perceber a melhor distribuição produzida pelo algoritmo **Com Filtro**. Por exemplo, ao se considerar o campo *etniaRedome* **Ignorado**, a distribuição para $N = 2$ e $N = 3$ foi exatamente igual a do SIVEP-Gripe (0.0), enquanto que para $N = 4$ foi muito similar (0.11). Por outro lado, quando analisamos os valores produzidos pelos outros dois algoritmos (**Sem Filtro** e **Aleatório**) se percebe um percentual muito maior.

Outra observação importante é que a distribuição produzida pelo algoritmo **Com Filtro** piora a medida que o valor de N aumenta. Esse comportamento é esperado

uma vez que, ao se aumentar o valor de N, a cada passo do algoritmo existirão menos opções de registros para serem usadas no próximo pareamento, o que pode fazer com que o algoritmo não consiga aplicar um ou mais filtros. Por exemplo, com $N = 4$ após se parear o primeiro registro do SIVEP-Gripe, 8 registros a menos estarão disponíveis (4 positivos e 4 negativos), enquanto que com $N = 2$ apenas 4 registros não estarão disponíveis. Esse mesmo comportamento não acontece nos outros dois algoritmos (**Sem Filtro** e **Aleatório**), pois a escolha dos registros não utiliza nenhum tipo de prioridade.

Além da frequência dos valores para cada raça/etnia, como esse campo é muito relevante para a análise da prevalência de alelos HLA, foi realizado o teste chi-quadrado para comparar duas variáveis categóricas e verificar se são homogêneas entre si. Neste caso, é comparado a variável *etniaRedome* do SIVEP-Gripe com a variável *etniaRedome* da base de controle e-SUS gerada pelo algoritmo. Os valores para **chi-2** na tabela mostram claramente que o algoritmo **Com Filtro** produziu os resultados mais homogêneos, uma vez que quanto menor o valor do **chi-2** melhor. Mais importante, a observação que a probabilidade ($\boldsymbol{p}$) do **chi-2** obtido foi 99.8% para $N = 2$ e 81.3% para $N = 3$ confirma a homogeneidade dos valores obtidos. Por sua vez, para $N = 4$, a probabilidade encontrada foi de 34.1%, mostrando que para esse valor de N a distribuição não está tão homogênea. Por outro lado, quando analisamos o valor de **chi-2** para os outros dois algoritmos (**Sem Filtro** e **Aleatório**), é possível observar que os resultados produzidos estão completamente desbalanceados, com probabilidade ($\boldsymbol{p}$) igual a 0 (zero).

3.1.3 Avaliação dos Campos Município e Região

O terceiro filtro aplicado é o do campo *municipio*, ou seja, o algoritmo de pareamento tenta escolher pessoas do mesmo município. Porém, em função da dificuldade de se conseguir na base de controle pessoas do mesmo município (o estado do Rio de Janeiro é composto de 92 municípios), o algoritmo aplica o filtro pelo campo *regiao* caso o filtro por município não retorne a quantidade mínima de N registros positivos e negativos desejada. Como dito anteriormente, o município/região é usado para a escolha de registros da base de controle que tenham sido infectados pela mesma variante da COVID-19. Outro motivo considerado para essa escolha é que os municípios pequenos podem apresentar uma distribuição alélica diferente da distribuição esperada. Exemplo: uma quantidade variável para cada tipo de raça/etnia por município. Em alguns, por exemplo, podem haver mais brancos ou então mais pretos e vice-versa.

Os municípios do estado do Rio de Janeiro são agrupados em 9 regiões: Baia da Ilha Grande, Baixada Litorânea, Centro-Sul, Médio Paraíba, Metropolitana I, Metropolitana II, Noroeste, Norte e Serrana. Os resultados da frequência (em porcentagem) para cada uma das regiões presentes nas bases de controle geradas pelos algoritmos podem ser

visualizados na Tabela 3.

Tabela 3 - Valores percentuais (%) para frequência dos valores para o campo *regiao*

Região	SIVEP	e-SUS								
		Com Filtro			Sem Filtro			Aleatório		
		2:1	3:1	4:1	2:1	3:1	4:1	2:1	3:1	4:1
I. Grande	3.29	2.17	2.17	2.27	2.30	2.60	2.64	3.64	2.66	2.60
Baix. Lit.	3.99	3.38	3.44	3.73	4.55	4.94	4.61	5.81	4.97	5.26
Cen. Sul	2.08	1.95	1.82	1.95	2.99	3.12	2.51	2.73	2.74	2.77
Med. Par.	6.76	8.84	9.21	9.64	12.09	12.05	10.38	12.78	12.94	12.69
Metrop. I	54.25	53.47	52.25	50.69	41.47	41.74	44.54	41.98	43.56	44.74
Metrop. II	10.92	13.08	13.98	14.34	18.02	16.78	18.76	13.21	13.78	13.63
Noroeste	0.17	0.48	0.52	0.89	3.12	2.60	2.66	2.12	3.00	2.38
Norte	2.95	2.77	3.09	3.21	4.72	5.03	4.44	5.24	4.65	4.31
Serrana	15.60	13.86	13.52	13.28	10.44	11.15	9.45	12.48	11.70	11.61

Apesar do campo *regiao* ser a quarta prioridade, o algoritmo de pareamento **Com Filtro** conseguiu uma distribuição balanceada de acordo com a frequência das regiões no SIVEP-Gripe. Por exemplo, se considerarmos a região **Metropolitana I** os valores percentuais para a frequência proporcionada pelo algoritmo **Com Filtro** são bem próximos da frequência da base de casos (SIVEP-Gripe). Por outro lado, a distribuição proporcionada pelos outros dois algoritmos (**Sem Filtro** e **Aleatório**) é claramente desbalanceada.

3.1.4 Avaliação dos Campos Sexo e Tipo de Teste

Com relação aos campos *sexo* (Tabela 4) e *tipoTeste* (Tabela 5) a distribuição gerada pelo algoritmo não foi tão homogênea. Apesar do balanceamento do algoritmo proposto ser muito superior aos outros dois algoritmos (**Sem Filtro** e **Aleatório**), quando comparada com a frequência da base de casos (SIVEP-Gripe) o pareamento ficou desbalanceado. Existem duas razões para esse comportamento. Primeiro, esses dois campos são os de menor prioridade. Segundo, e mais importante, a base e-SUS usada para gerar a base de controle é desbalanceada para esses campos em relação à base SIVEP-Gripe. Por exemplo, na base e-SUS 59.98% dos registros são do sexo feminino e 40.02% são do sexo masculino, enquanto que na base SIVEP-Gripe, 40.21% são do sexo feminino e 59.79% são do sexo masculino. A frequência para os valores para o campo *sexo* pode ser vista na Tabela 4.

No parágrafo anterior já foram descritos os motivos para o algoritmo proposto não conseguir balancear os registros para esse campo. Porém, mesmo o campo *sexo* tendo uma das últimas prioridades, a frequência gerada pelo algoritmo **Com Filtro** em relação

Tabela 4 - Valores percentuais (%) para frequência dos valores para o campo *sexo*

Sexo	**SIVEP**	**e-SUS**								
		Com Filtro			**Sem Filtro**			**Aleatório**		
		2:1	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**
Feminino	40.21	53.86	56.12	56.56	59.92	61.21	60.88	61.57	61.53	59.12
Masculino	59.79	46.14	43.88	43.44	40.08	38.79	39.12	38.43	38.47	40.88

a frequência gerada pelos outros dois algoritmos (**Sem Filtro** e **Aleatório**), foi superior. É importante destacar que o campo *sexo* não tem influência direta sobre os alelos HLA, de maneira que o desbalanceamento desse campo em relação a base de casos não prejudica a qualidade da base de controle para o estudo da prevalência de alelos HLA para pacientes com COVID-19.

Com relação ao campo *tipoTeste* não foi possível tentar parear com os registros do SIVEP-Gripe, pois esta base de dados não tinha essa informação. Assim, optou-se por selecionar diretamente registros do e-SUS dando-se prioridade para os testes **RT-PCR**, **Teste Rápido Antígeno** e **Anticorpo**, nessa ordem, conforme descrito na Subseção 2.3.1. A Tabela 5 apresenta a quantidade absoluta destes testes que cada algoritmo gerou.

Tabela 5 - Valores absolutos para o campo *tipoteste*

Tipo do teste	**e-SUS**								
	Com Filtro			**Sem Filtro**			**Aleatório**		
	2:1	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**
RT-PCR	1189	1686	2170	1024	1488	1876	995	1516	1990
TR Antígeno	121	187	253	185	240	297	334	505	689
TR Anticorpo	796	1279	1798	949	1458	2062	845	1281	1700

O algoritmo **ComFiltro** conseguiu gerar os arquivos com a maior quantidade de registros com *tipoTeste* **RT-PCR**, mesmo sendo esse campo o de menor prioridade do algoritmo. Como nesta tabela os valores são absolutos, então a quantidade de valores aumenta com o crescimento do valor de N. De maneira geral, o algoritmo proposto conseguiu atingir o objetivo de maximizar, dentro da sua prioridade, o tipo de teste RT-PCR.

3.2 Avaliação do Algoritmo de Pareamento MG

Esta seção avalia o algoritmo de pareamento utilizando as bases SIVEP-Gripe e e-SUS do estado de Minas gerais. Após todo o pré-processamento e busca dos alelos HLA no REDOME essas bases ficaram compostas com as seguintes quantidades de registros:

- o SIVEP-Gripe continha 1.479 registros;
- o e-SUS continha 31.625 registros.

A quantidade de registros do arquivo final correspondente a base de controle após o pareamento depende do valor do N escolhido. Por exemplo, para $N = 4$ o arquivo de casos terá $2 \times 4 \times 1.479 = 11.832$ registros, onde 5.916 registros são positivos para COVID-19 e 5.916 são negativos.

As subseções a seguir apresentam a avaliação da homogeneidade para cada um dos campos selecionados. Além disso, é importante destacar, novamente, que a avaliação foi realizada para 3 valores de N distintos: 2, 3 e 4.

3.2.1 Avaliação do Campo Idade

Inicialmente, foi avaliada a distribuição do campo *idade*. A mesma divisão por faixa etária utilizada para a base do Rio de Janeiro (Subseção 2.3.2) foi realizada para base de Minas Gerais. A menor idade, maior idade e idade média para cada uma das bases de controle geradas pelos algoritmos podem ser vistas na Tabela 6.

Tabela 6 - Valores máximo, médio e mínimo para o campo *idade*

Idade	**SIVEP**	**e-SUS**								
		Com Filtro			**Sem Filtro**			**Aleatório**		
		2:1	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**
Mínima	20	20	20	20	1	1	1	1	1	1
Média	47	45	45	43	38	38	38	38	38	38
Máxima	78	78	78	78	76	76	76	79	76	79

Assim como ocorreu para a base do Rio de Janeiro (Subseção 3.1.1), o algoritmo de pareamento proposto (**Com Filtro**), proporcionou uma escolha muito mais homogênea. Por exemplo, enquanto que a menor idade escolhida foi 20 (igual a do SIVEP-Gripe), a dos outros dois algoritmos (**Sem Filtro** e **Aleatório**) foi 1 (um). Além disso, a média das idades (45) ficou muito mais próxima da média do SIVEP-Gripe (47), do que a média dos outros dois algoritmos (38). Com relação a variação do valor de N, para $N = 4$ a média das idades ficou ligeiramente menor (43). A explicação para isso é a menor quantidade de opções a serem escolhidas quando se aumenta o valor de N.

3.2.2 Avaliação do Campo Etnia

Assim como foi realizado para a base do RJ, por ser o campo *etniaRedome* o mais importante do algoritmo de pareamento proposto, uma vez que essa característica tem uma influência significativa no alelo que pode ser encontrado, ele tem a maior prioridade do algoritmo (NUNES et al., 2020). Os resultados da frequência (em porcentagem) para cada uma das etnias presentes nas bases de controle geradas pelos algoritmos, assim como o valor do chi-quadrado, podem ser visualizados na Tabela 7.

Tabela 7 - Valores em percentual (%) para frequência dos valores para o campo *etniaRedome*

Etnia	**SIVEP**	**e-SUS**								
		Com Filtro			**Sem Filtro**			**Aleatório**		
		2:1	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**
Amarela	1.69	1.69	1.75	1.76	1.61	1.65	1.77	1.83	1.75	1.57
Branca	39.69	39.69	39.90	39.85	42.97	43.25	43.42	43.49	43.99	42.56
Indígena	0.47	0.42	0.33	0.30	0.20	0.21	0.22	0.20	0.17	0.22
Não Inf.	15.89	15.92	15.67	15.24	11.2	10.67	10.76	11.07	10.83	11.38
Parda	31.24	31.25	31.47	31.86	35.11	34.87	34.48	34.38	33.83	34.98
Preta	11.02	11.02	10.88	11.00	8.89	9.35	9.36	9.03	9.43	9.28
chi-2		0.07	1.05	2.14	38.53	45.93	46.11	38.76	44.91	38.08
p		99.9%	95.8%	82.8%	0%	0%	0%	0%	0%	0%

Comparando os valores de frequência para cada uma das etnias das bases de controle (e-SUS) geradas pelos algoritmos de pareamento com os valores da base de casos (SIVEP-Gripe) pode-se perceber claramente a melhor distribuição produzida pelo algoritmo proposto (**Com Filtro**). Por exemplo, ao se considerar a *etniaRedome* **Branca**, a distribuição para $N = 2$, $N = 3$ e $N = 4$ foram, respectivamente, 39.69, 39.90 e 39.85, muito próxima ou até mesmo igual ($N = 2$) que a distribuição do SIVEP-Gripe. Por outro lado, quando se analisa os valores produzidos pelos outros dois algoritmos (**Sem Filtro** e **Aleatório**) se percebe um percentual muito maior.

Outra observação relevante, é que a distribuição produzida pelo algoritmo **Com Filtro** tende a piorar a medida que o valor de N aumenta. Esse comportamento é esperado uma vez que, ao se aumentar valor de N, a cada passo do algoritmo existirão menos opções de registros para serem usadas no próximo pareamento, o que pode fazer com que o algoritmo não consiga aplicar um ou mais filtros. Esse mesmo comportamento não acontece nos outros dois algoritmos (**Sem Filtro** e **Aleatório**), pois a escolha dos registros não utiliza nenhum tipo de prioridade.

Além da frequência dos valores para cada *etniaRedome*, como esse campo é muito relevante para a análise da prevalência de alelos HLA, foi realizado o teste chi-quadrado para comparar duas variáveis categóricas e verificar se são homogêneas entre si. Os valores

para **chi-2** apresentados na última linha da Tabela 7 mostram claramente que o algoritmo **Com Filtro** produziu os resultados mais homogêneos, uma vez que quanto menor o valor do **chi-2** melhor. Mais importante, a observação que a probabilidade do **chi-2** obtido foi 99.9% para $N = 2$, 95.8% para $N = 3$ e 82.8% para $N = 4$, o que confirma a homogeneidade dos valores obtidos. Por outro lado, quando analisamos o valor de **chi-2** para os outros dois algoritmos (**Sem Filtro** e **Aleatório**), é possível observar que os resultados produzidos são claramente desbalanceados, com valor de probabilidade igual 0%.

3.2.3 Avaliação do Campo Sexo

Com relação ao campo *sexo* (Tabela 8), o pareamento gerado pelo algoritmo proposto ficou desbalanceado, ainda que o resultado seja muito superior aos outros dois algoritmos (**Sem Filtro** e **Aleatório**), quando comparada com a frequência da base de casos (SIVEP-Gripe). As razões para o desbalanceamento são duas. Primeiro. o campo *sexo* é um campo de menor prioridade, ou seja, os campos *idade* e *etniaRedome*, por exemplo têm maior prioridade. Segundo, da mesma forma que na base do RJ, a base e-SUS usada para gerar a base de controle é desbalanceada para esse campo em relação à base SIVEP-Gripe. Enquanto que na base e-SUS 59.09% dos registros são do sexo feminino e 40.91% são do sexo masculino, na base SIVEP-Gripe, 43.48% são do sexo feminino e 56.51% são do sexo masculino. A frequência para os valores para o campo *sexo* pode ser vista na Tabela 8.

Tabela 8 - Valores percentuais (%) para frequência dos valores para o campo *sexo*

Sexo	**SIVEP**	**e-SUS**								
		Com Filtro			**Sem Filtro**			**Aleatório**		
		2:1	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**	**2:1**	**3:1**	**4:1**
Feminino	43.48	53.45	55.28	55.88	61.05	60.42	60.66	59.69	59.79	60.03
Masculino	56.51	46.18	44.33	43.69	38.52	39.17	39.00	39.89	39.87	39.63

Similarmente ao que ocorreu na base do RJ, a frequência gerada pelo algoritmo **Com Filtro** em relação a frequência gerada pelos outros dois algoritmos (**Sem Filtro** e **Aleatório**) foi superior ainda que o campo *sexo* tenha uma das últimas prioridades. Vale destacar novamente que o campo *sexo* não tem influência direta sobre os alelos HLA. Assim, o desbalanceamento desse campo em relação a base de casos não prejudica a qualidade da base de controle para o estudo da prevalência de alelos HLA para pacientes com COVID-19.

Assim como para a base RJ, o algoritmo proposto também tentou parear o campo

município/região. Porém, devido a grande quantidade de municípios e também de regiões, o pareamento para este campo ficou muito desbalanceado. Com relação ao campo $tipoTeste$, não foi possível priorizar o teste **RT-PCR** pois o arquivo e-SUS de MG não continha dados para este campo.

3.3 Análise da Prevalência de Alelos HLA MG

A análise da prevalência de alelos HLA foi realizada utilizando os dados dos arquivos SIVEP-Gripe e e-SUS de pacientes residentes no estado de Minas Gerais. Após todas as etapas descritas na Seção 2.2, a base de casos SIVEP-Gripe ficou composta de 1479 registros e a base e-SUS para geração da base de controle com 31625 registros. Foi utilizado o arquivo de pareamento com $N = 2$ gerado pelo algoritmo de pareamento proposto. O arquivo gerado é balanceado, especialmente com relação à etnia, e o pareamento ficou o mais próximo possível dentro das opções dos registros disponíveis. Foi realizada a análise que comprovou diferenças entre a distribuição dos grupos alélicos em função da autodeclaração raça/etnia. Mais especificamente, não foram encontradas diferenças entre a distribuição dos grupos alélicos nos locos A e DRB1 em pacientes com raça/etnia **Branca**, nos locos A, B e DRB1 em pacientes com raça/etnia **Parda** e nos locos B e DRB1 com raça/etnia **Preta**. Por sua vez, quando comparados os 4 grupos de acordo com a evolução do estado do paciente (**Recuperado**, **Internado**, **UTI** e **Óbito**) existem diferenças na frequência dos grupos alélicos do loco B em pacientes com raça/etnia **Branca** e do loco A em pacientes com raça/etnia **Preta**. A comparação entre a frequência alélica do Alelo A*36 foi maior nos pacientes autodeclarados com raça/etnia **Preta** internados com COVID-19 em UTI que no grupo **Recuperado** (razão de chance $9,6$ IC: $2,6 - 34,8$; $p = 0,001$) ou no Grupo **Internado** (razão de chance 7,8 IC: 1,9-32,5; p=0,005), indicando que este alelo oferece um risco maior para pacientes com COVID-19. Por outro lado, a comparação entre a frequência alélica do alelo B*51 foi significativamente maior nos pacientes autodeclarados com raça/etnia **Branca** sem internação com COVID-19 do que nos Internados (razão de chance $0,7$ IC: $0,5 - 0,9$; $p = 0,021$) ou do Grupo **Óbito** (razão de chance $0,5$ IC: $0,2 - 0,9$; $p = 0,014$), indicando que este alelo oferece uma proteção maior a COVID-19.

CONCLUSÃO

A saúde pode ser considerada como um setor da economia responsável por bens e serviços que visam pela qualidade de vida dos indivíduos sendo que no Brasil tem-se o SUS como o sistema de saúde público que coordena as ações das esferas de governo. Assim como outros setores, a saúde passou por processos de transformação digital através da utilização dos SIS que provêm mecanismos para auxiliar profissionais e órgãos de saúde no processo de tomada de decisões. No âmbito do SUS, vários SIS foram usados para auxiliar no combate à pandemia da COVID-19. Entre eles podem ser destacados o SIVEP-Gripe e o e-SUS ambos usados para monitorar casos de COVID-19.

A proposta deste trabalho foi um ECC no qual um algoritmo de pareamento gera uma base de controle com registros do e-SUS selecionados conforme algumas características determinadas a partir de registros do SIVEP-Gripe. O ECC abordado avalia a prevalência de alelos HLA em pacientes com COVID-19 e propõe tal algoritmo de pareamento capaz de criar uma base de controle homogênea (e-SUS), a partir das características dos registros da base de casos (SIVEP-Gripe). O SIVEP-Gripe é usado para gerar a base de casos com registros de notificações de pacientes com casos mais graves ou de óbito de COVID-19. O e-SUS é usado para gerar a base de controles com registros de notificações de pacientes com casos suspeitos ou confirmados de COVID-19. O objetivo final é ter uma base consistente da qual se possa realizar estudos sobre a presença de genes alelos em pacientes que tiveram a COVID-19. Deseja-se extrair informações úteis que auxiliem no processo de identificação de prevalência de genes alelos em pacientes diagnosticados com COVID-19.

Os registros foram coletados por usuário autenticado e com acesso aos sistemas SIVEP-Gripe e e-SUS e disponibilizados em arquivos. Os dados passaram por um processo de pré-processamento envolvendo uma uniformização, limpeza e integração para remover inconsistências, duplicidades e outros tipos de anomalias encontradas que pudessem dificultar no uso destas informações na identificação de genes alelos. Após isso, os dados foram anonimizados conforme LGPD e persistidos em um banco de dados relacional.

Como era essencial ter as informações sobre os alelos dos pacientes e visto que a tipificação HLA (Silva, Marcio NP and Cristóvão, Luís and Pôrto, MS and Marzulo, Leandro AJ and Sena, Alexandre C, 2019) é um processo caro, os registros persistidos no banco de dados foram consultados da base do REDOME.

Para que os dados do ECC proposto fosse o mais homogêneo possível, este trabalho propôs um algoritmo de pareamento responsável por gerar uma base de controle com registros do e-SUS pareados com registros do SIVEP-Gripe aplicando filtros sobre determinados campos como: a etnia/raça, o sexo, a localidade (por município ou região) e o tipo de teste realizado. Os filtros são realizados em cascata e a cada iteração é avaliada

a necessidade de remoção de algum filtro quando a quantidade mínima de registros a ser pareada não é alcançada. O número de registros a serem pareados é um parâmetro de entrada definido pelo algoritmo que estabelece a quantidade de registros do e-SUS com resultado de testes positivos e negativos. A ideia é que a proporção entre registros e-SUS e SIVEP-Gripe seja a mais balanceada possível com relação às características escolhidas para o pareamento.

Com isso, foi possível ter uma base de informações confiável, sobre pacientes que tiveram COVID-19 incluindo informações sobre os seus alelos HLA. Isso pode ajudar aos especialistas em imunologia não só na identificação de genes alelos que ajudem na proteção da COVID-19 como também em geles alelos que indiquem maiores riscos de complicações por esta doença. Desse modo, o algoritmo poderá ajudar a comunidade científica provendo uma base de controle balanceada em relação aos casos a serem estudados respeitando a ordem de prioridade das características escolhidas. Além disso, o uso deste algoritmo de pareamento isenta o próprio pesquisador ter de escolher os casos da base de controle, evitando assim um possível viés de seleção que é um dos principais problemas da técnica de ECC.

A análise do algoritmo de pareamento evidenciou a escolha balanceada do pareamento considerando os campos com maior prioridade para realizar os filtros. Ao se comparar o algoritmo proposto com outros dois algoritmos (**Sem Filtro** e **Aleatório**) que não utilizam nenhum tipo de prioridade, mostra-se os benefícios do pareamento alcançado pelo mesmo. Por fim, a análise da prevalência de alelos HLA mostrou que existem diferenças na distribuição dos grupos alélicos em função do campo raça/etnia, onde o alelo B*51 tem uma chance maior de oferecer proteção, enquanto que o alelo A*36 aumenta o risco para a COVID-19.

Como trabalhos futuros, será realizada a análise da prevalência de alelos HLA para a base do estado do Rio de Janeiro. Já foram recebidos novos dados, de maneira que será possível conseguir mais de 1000 pacientes para a base de casos, o que é a quantidade mínima necessária para se fazer um estudo de alelos confiável. Além disso, o algoritmo de pareamento será aprimorado para que ele possa ser utilizado para qualquer estudo de caso, bastando o usuário selecionar os campos.

REFERÊNCIAS

Bahia, Governo do Estado da. *Guia Rápido SIVEP GRIPE*. Accessed: 2022-01-15. Disponível em: ⟨http://www.saude.ba.gov.br/wp-content/uploads/2021/05/GUIA-RAPIDO-SIVEP-GRIPE-atualizado-em-maio_2021.pdf⟩.

Brasil, Ministério da Sáude. *e-SUS Notifica*. Accessed: 2022-01-08. Disponível em: ⟨{https://www.gov.br/saude/pt-br/composicao/svs/sistemas-de-informacao/e-sus-notific}.⟩

_____. *Orientações sobre notificação e registros de casos de Covid-19 no Brasil*. Accessed: 2022-01-08. Disponível em: ⟨{https://www.gov.br/saude/pt-br/coronavirus/artigos/notificacao-e-registr}.⟩

_____. *Plano Diretor de Tecnologia da Informação e Comunicação - 2019 — 2021*. Accessed: 2023-03-03. Disponível em: ⟨{https://datasus.saude.gov.br/wp-content/uploads/2020/05/22052020v5.pdf}.⟩

_____. *Portaria Conjunta SE/MS/SAS n o 23 de 21/05/2004 - Federal - LegisWeb [Internet]*. Accessed: 2023-03-03. Disponível em: ⟨{https://www.legisweb.com.br/legislacao/?id=189755}.⟩

_____. *PORTARIA No 589, DE 20 DE MAIO DE 2015, Política Nacional de Informação e Informática em Saúde (PNIIS)* . Accessed: 2023-03-03. Disponível em: ⟨{https://bvsms.saude.gov.br/bvs/saudelegis/gm/2015/prt0589_20_05_2015.html}.⟩

_____. *Registro Brasileiro de Doadores Voluntários de Medula Óssea*. Accessed: 2023-03-03. Disponível em: ⟨{https://redome.inca.gov.br/profissional-de-saude/registro-nacional-de-receptores-de-medula-ossea-rereme}.⟩

_____. *SIVEP-Gripe*. Accessed: 2022-01-08. Disponível em: ⟨{https://sivepgripe.saude.gov.br/sivepgripe}.⟩

_____. *Sobre o DATASUS*. Accessed: 2022-01-08. Disponível em: ⟨https://datasus.saude.gov.br/sobre-o-datasus/⟩.

Cavalcante, Ricardo Bezerra and Ferreira, Marina Nagata and Silva, Poliana Cavalcante. Sistemas de informação em saúde: possibilidades e desafios. *Revista de Enfermagem da UFSM*, v. 1, n. 2, p. 290–299, 2011.

CORREALE, P. et al. Hla-b*44 and c*01 prevalence correlates with covid19 spreading across italy. *International journal of molecular sciences*, v. 21, 2020.

Correale P, Mutti L, Pentimalli F, et al. Hla-b*44 and c*01 prevalence correlates with covid19 spreading across italy. *Int J Mol Sci*, v. 21, n. 15, 2020.

de Fátima Marin, Heimar. Sistemas de informação em saúde: considerações gerais. *Journal of Health Informatics*, v. 2, n. 1, 2010.

Epi Info. *Epi Info*. Accessed: 2022-03-04. Disponível em: ⟨{https://www.cdc.gov/epiinfo/por/pt_index.htm}.⟩

Federal, Senado. Constituição (1988). Constituição da República Federativa do Brasil. *Brasília (DF)*, 1988.

Freitas, Carlos Machado de and Barcellos, Christovam and Villela, Daniel Antunes Maciel. Covid-19 no brasil: cenários epidemiológicos e vigilância em saúde. Série Informação para ação na Covid-19— Fiocruz, 2021.

LOPES, Claudia S; COUTINHO, Evandro SF. Transtornos mentais como fatores de risco para o desenvolvimento de abuso/dependência de cocaína: estudo caso-controle. *Revista de Saúde Pública*, SciELO Brasil, v. 33, n. 5, p. 477–486, 1999.

MACHADO, Luiz Gustavo et al. Infecções relacionadas à assistência à saúde no brasil: Prevalência multicêntrica e estudo caso-controle pareado. *The Brazilian Journal of Infectious Diseases*, Elsevier, v. 26, p. 102252, 2022.

MENDES, Gabriel P. et al. Análise da prevalência de alelos hla em pacientes com covid-19. *Congresso Brasileiro de Informática em Saúde*, SBIS, 2022. ISSN 0000-0000.

NUNES, Kelly et al. How ancestry influences the chances of finding unrelated donors: An investigation in admixed brazilians. *Frontiers in Immunology*, v. 11, 2020. ISSN 1664-3224. Disponível em: ⟨https://www.frontiersin.org/articles/10.3389/fimmu.2020.584950⟩.

OLIVEIRA, Marco Aurelio; VELLARDE, Guillermo Coca; SÁ, Renato Augusto Moreira de. Entendendo a pesquisa clínica iv: estudos de caso controle. *Femina*, p. 175–180, 2015.

Paim, Jairnilson Silva. *O que é o SUS*. [S.l.]: SciELO-Editora FIOCRUZ, 2009.

_____. Sistema único de saúde (sus) aos 30 anos. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 23, p. 1723–1728, 2018.

PEREIRA, Susan Martins et al. Associação entre diabetes e tuberculose: estudo caso controle. *Revista de Saúde Pública*, SciELO Brasil, v. 50, 2016.

PINTO, Luiz Felipe; FREITAS, Marcos Paulo Soares de; FIGUEIREDO, André William Sant'Anna de. Sistemas nacionais de informação e levantamentos populacionais: algumas contribuições do ministério da saúde e do ibge para a análise das capitais brasileiras nos últimos 30 anos. *Ciência Saúde Coletiva*, ABRASCO - Associação Brasileira de Saúde Coletiva, v. 23, n. Ciênc. saúde coletiva, 2018 23(6), p. 1859–1870, Jun 2018. ISSN 1413-8123. Disponível em: ⟨https://doi.org/10.1590/1413-81232018236.05072018⟩.

RÊGO, Marco Antônio V. Estudos caso-controle: uma breve revisão. *Gazeta Médica da Bahia*, n. 1, 2010.

S KERNDT CC, Hoffman MR Tenny. Case control studies. In: *StatPearls [Internet]*. [S.l.]: StatPearls Publishing, 2022.

Santos, Alethele de Oliveira and Lopes, Luciana Tolêdo. Planejamento e gestão. In: *Planejamento e gestão*. [S.l.: s.n.], 2021. p. 342–342.

Saúde, Ministério da. *Tutorial de Navegação Versão 5 Agosto de 2021*. Accessed: 2022-02-12. Disponível em: ⟨https://datasus.saude.gov.br/wp-content/uploads/2021/08/Tutorial-de-Navegacao-e-SUS-VE_16_08_21.pdf⟩.

Silva, Marcio NP and Cristóvão, Luís and Pôrto, MS and Marzulo, Leandro AJ and Sena, Alexandre C. Estudo e implementação de um sistema customizável para controle laboratorial para o processo de tipificação hla. In: SBC. *Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*. [S.l.], 2019. p. 118–129.

STRALEN, Karlijn J Van et al. Case-control studies–an efficient observational study design. *Nephron Clinical Practice*, Karger Publishers, v. 114, n. 1, p. c1–c4, 2010.

APÊNDICE

Vale destacar que os exemplos mostrados nas explicações contêm *dados fictícios* não sendo exemplos reais retirados de algumas das bases originais do SIVEP-Gripe e e-SUS, zelando pelo sigilo das informações conforme a LGPD.

Uniformização dos Dados

Sendo um cadastro digitalizado e dependente de usuários como profissionais e unidades de saúde, erros originários da digitação ou tipificação dos dados podem surgir. Campos como nome não possuem padrão e, assim, é possível cadastrar duas notificações para um mesmo paciente com alguma variação na tipografia dos dados. Por exemplo: se existe uma paciente fictícia chamada "MARIA DAS GRAÇAS" é possível cadastrá-la ainda como "Maria Das Gracas", "Maria Das Graças", "Maria Das Gracas" ou ainda "MARIA Das Gracas". Os principais problemas que podem surgir para identificar que se trata do mesmo paciente é perceber as possíveis divergências na capitalização (letras maiúsculas e minúsculas), acentuação ou presença de espaços em branco.

Todas estas questões foram levadas em consideração para tratar estes dados já que era essencial produzir as bases já com os dados consolidados e aptos para serem utilizados nas etapas posteriores. Primeiramente, a ideia foi padronizar todos os campos possíveis para poder então remover a maioria das duplicidades e gerar registros únicos para cada paciente.

Com relação à tipografia dos dados nas bases SIVEP-Gripe e e-SUS (tanto do estado do Rio de Janeiro quanto do estado de Minas Gerais), observaram-se os seguintes problemas:

- falta de padronização de campos entre as bases SIVEP-Gripe e e-SUS:
  - campo *sexo* com valores "M" e "F" ou com os valores "Masculino" e "Feminino";
  - campo *cpf* informado apenas com números "XXXXXXXXXXX" ou formatado como *string* "XXX.XXX.XXX-XX";
  - campo *dataInternacao* apresentava o formato "yyyy-MM-dd hh:mm:ss" enquanto as outras datas apresentavam o formato "dd/MM/yyyy".
- presença de espaços em branco (em qualquer posição):
  - campo *nomeCompleto* apresentando valores como "MARIA SANTOS" ou ainda " MARIA SANTOS ";

- presença de acentos e caracteres especiais:
  - campo *nomeCompleto* apresentando valores como “EDNEIA DA SILVA” ou ainda “EDNÉIA DA SILVA”.

Para padronizar os campos foi necessário aplicar a mesma formatação para persistí-los no banco de dados sendo alguns exemplos:

- capitalização de caracteres para letras maiúsculas em campos *string* como, por exemplo, o *nomeCompleto*;
- campos do tipo data formatados para um mesmo padrão;

Outra questão tratada foi que em alguns dos arquivos *.csv* referentes a uma das bases, haviam registros possuindo o “;” como parte do seu conteúdo mas este mesmo caractere era utilizado como separador entre os valores dos campos. Isso gerava um problema quando era realizada a leitura destes registros e percebeu-se que um mesmo carácter não poderia ser utilizado como separador entre os campos se também fizesse parte do conteúdo dos mesmos. Um exemplo observado foi com o campo *descricaoOutros* sobre o qual havia um registro na base com o valor “MIALGIA; ODINOFAGIA”. Quando esse campo era analisado, interpretavam-se dois valores diferentes: “MIALGIA;” e “ODINOFAGIA” sendo que na verdade tratava-se de apenas uma única informação. Assim, foi necessário aplicar uma pequena formatação substituindo nestes arquivos o “;” pelo caractere “,”. Dessa forma, quando fosse feita a leitura do campo *descricaoOutros* seria considerado todo o valor “MIALGIA; ODINOFAGIA” já que antes da formatação apenas a parte “MIALGIA” era considerada.

Limpeza dos Dados

O objetivo da uniformização dos dados foi de colocá-los com uma padronização e formatação únicas para realizar em seguida uma limpeza sobre eles. Esta limpeza consistiu na identificação de duplicidades e inconsistências encontradas entre os registros que necessitavam ser removidas para assim então persistir as informações no banco de dados.

As duplicidades consistiam em registros de notificações que pertencessem ao mesmo paciente. Todas as notificações foram persistidas no banco de dados. Contudo, desejava-se também identificar e eliminar as redundâncias entre as notificações afim de também *manter no banco de dados um único registro por paciente.* Tanto as notificações quanto os registros únicos de pacientes gerados foram mantidos em tabelas separadas no banco de dados.

Para tentar identificar as notificações referentes ao mesmo paciente, analisou-se os campos *nomeCompleto*, *cpf* e *dataNascimento*. O critério adotado foi considerar a combinação dos valores destes campos como uma espécie de *chave* para os registros de notificações. Todavia, alguns problemas foram encontrados durante esta etapa:

- notificações contendo o campo *nomeCompleto* informado mas com *cpf* em branco;
- notificações contendo o campo *cpf* informado mas com *nomeCompleto* em branco;

O único campo presente entre todos os registros era a *dataNascimento*.

Assim, foi necessário complementar tais informações entre os registros de notificações referentes ao mesmo paciente preenchendo o *nomeCompleto* ou o *cpf*. Então a estratégia foi buscar por notificações que tivessem os três campos informados para então capturar seus valores e inserir os dados ausentes de *nomeCompleto* ou *cpf* nas outras notificações que tivessem um desses campos ausentes. Essa abordagem não capturou todas as possíveis duplicidades já que nem todas as notificações referentes ao mesmo paciente continham todos os campos preenchidos. Algumas outras duplicidades foram identificadas e removidas manualmente apenas posteriormente e outras foram removidas baseando-se na combinação de valores do campo *nomeCompleto* com outros campos como a *dataNotificacao*, a *classificacaoFinal*, a *evolucaoCaso*, a *dataInternacao* e a *dataEncerramento*.

No geral, tentou-se manter como um registro único para cada paciente aquelas notificações que tivessem as informações mais completas ou assertivas possíveis. Infelizmente, essa foi uma operação complicada de ser automatizada já que além do *nomeCompleto* ou do *cpf* haviam também outros campos ausentes entre alguns registros o que dificultou na identificação de todas as possíveis duplicidades. Além disso, existiam registros de notificações referentes ao mesmo paciente (com os mesmos valores para os campos *nomeCompleto*, *cpf* ou *dataNascimento*) que apresentavam divergência em campos como *logradouro* ou *municipio*. Ou seja, haviam notificações referentes ao mesmo paciente com endereços distintos. Nestes casos, não teve outra solução a não ser escolher aleatoriamente apenas uma das notificações como sendo o registro a ser persistido na tabela *paciente*. Por conta destas e de outras possíveis inconsistências que possam não ter sido capturadas (tanto de forma automatizada quanto de maneira manual) não pôde-se garantir a identificação e remoção de todas as duplicidades presentes nas bases SIVEP-Gripe e e-SUS.

Além da remoção das duplicidades alguns outros ajustes foram observados como sendo necessários para uso posterior pelo algoritmo de pareamento (ver Seção 2.3). Um exemplo é relacionado com o campo *evolucaoCaso* que informa o grau de evolução da COVID-19 para cada paciente. Supondo que um mesmo paciente tenha duas notificações com o campo *evolucaoCaso* diferentes sendo uma como "OBITO" e outra como "INTERNADO". Para o algoritmo de pareamento o interesse maior é em saber que o paciente

veio a óbito ainda que em outro momento ele tenha sido apenas internado. Ou seja, a preferência é usar a notificação para gerar o registro único de cada paciente que apresente *o estado mais grave* que ele já notificou. Considerou-se como *grau* para o campo *evolucaoCaso* a ordem a seguir (da mais alta para a mais baixa): "OBITO", "UTI', "INTERNADO", e "RECUPERADO". O campo *evolucaoCaso* categoriza os registros de acordo com a evolução do estado do paciente. Os pacientes com *evolucaoCaso* = "RECUPERADO" não precisaram ser internados e conseguiram se recuperar da doença. Por sua vez, os pacientes com *evolucaoCaso* = "INTERNADO" tiveram que ser internados, mas não precisaram de UTI e conseguiram se recuperar da doença. Já, os pacientes com evolução do caso *evolucaoCaso* = "UTI" tiveram que ser internados em UTIs, mas conseguiram se recuperar da doença. Por fim, os pacientes com *evolucaoCaso* = "OBITO" morreram de COVID-19.

Uma outra questão tratada relacionava-se com o campo $resultadoTeste$. Haviam registros de notificações para um mesmo paciente apresentando valores diferentes neste campo. No contexto da pandemia da COVID-19, o interesse é em saber se o paciente já teve a doença e assim, o que prevaleceu foi em utilizar as notificações como registro de paciente que apresentassem para o campo $resultadoTeste$ os seguintes valores em ordem de preferência: "POSITIVO", seguido por "NEGATIVO" e por último "INCONCLUSIVO".

Por último, também foi necessário identificar registros de notificações referentes ao mesmo paciente que por ventura estivessem presentes em ambas as bases SIVEP-Gripe e e-SUS. A ideia era formar os grupos de casos (SIVEP-Gripe) e de controle (e-SUS) garantindo que não houvessem notificações de um mesmo paciente em ambas as bases. Cada registro de notificação utilizado para representar exclusivamente um único paciente só deveria pertencer unicamente a um dos grupos: caso ou controle. Tal restrição era necessária para a execução do algoritmo de pareamento visando-se manter um balanceamento entre os dados presentes em ambas as bases. Usado como grupo de controle, o e-SUS servirá como base de comparação para o grupo de análise SIVEP-Gripe que poderá ser testado posteriormente, por exemplo, para verificar os genes alelos mais presentes em decorrência da gravidade (campo *evolucaoCaso*) dos caso de COVID-19.

Persistência dos Dados

Após a remoção de anomalias identificadas entre os dados (duplicidades e inconsistências) na etapa de limpeza, todos os registros foram salvos em duas tabelas no banco de dados:

- *notificacao*: armazena todos os registros de notificações presentes nos arquivos *.csv* originais passados pelo usuário autenticado aos sistemas SIVEP-Gripe e e-SUS. Além de todos os campos presentes nos arquivos *.csv*, foi acrescentado um campo

artificial nesta tabela chamado de *descartada* que indicava se tal registro de notificação poderia ser candidato a virar o registro único de notificação de um paciente;

- *paciente*: armazena registros únicos de notificações por paciente. Levou-se em consideração as notificações que apresentassem o número de informações mais completas possíveis visando principalmente o preenchimento dos campos *nomeCompleto*, *cpf* e *dataNascimento* usados para identificar unicamente cada paciente.

Ao final, os dados presentes em ambas as bases estavam persistidos e integrados ao banco de dados relacional e estavam aptos para serem utilizados posteriormente nas etapas futuras até serem utilizados pelo algoritmo de pareamento.

Ambas as tabelas mantém todos os campos já citados anteriormente em exemplos anteriores além de vários outros campos presentes nos arquivos *.csv* sendo que nem todos valem a pena ser mencionados por não serem utilizados. Vale destacar os principais campos que são levados em consideração pelo algoritmo de pareamento sendo eles:

- *nomeCompleto*;
- *cpf*;
- *dataNascimento*;
- *dataNotificacao*;
- *dataInicioSintomas*;
- *resultadoTeste* que pode apresentar os valores: “POSITIVO”, “NEGATIVO” ou “INCONCLUSIVO”;
- *dataTeste*;
- *tipoTeste* que pode apresentar os valores: “TESTE RÁPIDO - ANTICORPO”, “RT-PCR”, “TESTE RÁPIDO - ANTÍGENO”, “Imunoensaio por Eletroquimioluminescência - ECLIA IgG”, “Imunoensaio por Eletroquimioluminescência – ECLIA”, “Enzimaimunoensaio - ELISA IgM”, “Quimioluminescência - CLIA” e “Enzimaimunoensaio – ELISA”;
- *dataEncerramento*;
- *evolucaoCaso* que pode apresentar os valores: “OBITO”, “UTI”, “INTERNADO” e “RECUPERADO”;

Além destes campos, também foi gerado um campo artificial chamado *idade* derivado dos campos *dataNascimento* e *dataNotificacao*. Esse campo é usado pelo algoritmo de pareamento para agrupar os pacientes por faixas etárias sendo mais detalhado no Seção 2.3.

Tecnologias Empregadas

Optou-se por utilizar a linguagem de programação Java pois ela possui bibliotecas robustas para se trabalhar com a manipulação de arquivos e também com conexão para bancos de dados relacionais. Para ler ou editar os arquivos CSV (Comma-separated values) utilizou-se a biblioteca Open CSV que permite realizar estas operações de forma fácil e bem flexível. Dentre os principais recursos desta biblioteca pode-se listar a possibilidade de realizar a análise (*parse*) dos dados para diversos formatos, facilitando a leitura ou a escrita dos arquivos. Com o Open CSV é possível capturar os dados em formato *.csv* e guardá-los em atributos definidos em classes Java. O banco de dados relacional usado para criar as tabelas foi o MySQL e utilizou-se o framework JPA (acrônimo para Java Persistence API) que trabalha com o conceito de ORM (do inglês Object-Relational Mapping), ou seja, o mapeamento objeto-relacional. O JPA facilita a transcrição de entidades implementadas por classes Java para tabelas no banco de dados e isenta o programador de ter de lidar diretamente com código SQL nativo para realizar operações comuns como buscas, inserções, atualizações ou remoções sobre os dados.