



PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS

ESCOLA DE CIÊNCIAS MÉDICAS E DA VIDA

**FATORES DE RISCO PERINATAIS PARA O DIAGNÓSTICO PRECOCE DO
TRANSTORNO DO ESPECTRO AUTISTA EM GOIÁS: UM ESTUDO CASO-
CONTROLE COM LINKAGE DE BASES DE DADOS PÚBLICAS**

NAYÃ PEREIRA PORTILHO

GOIÂNIA – GO

2026

NAYÃ PEREIRA PORTILHO

TRABALHO DE CONCLUSÃO DE CURSO

**FATORES DE RISCO PERINATAIS PARA O DIAGNÓSTICO PRECOCE DO
TRANSTORNO DO ESPECTRO AUTISTA EM GOIÁS: UM ESTUDO CASO-
CONTROLE COM LINKAGE DE BASES DE DADOS PÚBLICAS**

Trabalho de Conclusão de Curso apresentado à Escola de Ciências Médicas e da Vida, da Pontifícia Universidade Católica de Goiás, como requisito parcial para obtenção do título de bacharel em Medicina.

Orientador: Prof. Dra. Graziela Torres Blanch

GOIÂNIA – GO

2026

AGRADECIMENTOS

Acima de tudo, agradeço a Deus. Sua presença constante em minha vida foi o alicerce que me sustentou nos momentos de dúvida e cansaço. Foi nos silêncios das noites de estudo e nas preces silenciosas antes de cada desafio que encontrei forças para seguir em frente. A Ti, Senhor, entrego este trabalho como fruto da minha dedicação, mas, sobretudo, da Tua infinita bondade e misericórdia.

Aos meus pais, Salvador Gomes Portilho e Sandra Lúcia Pereira Portilho, meu porto seguro e minha maior inspiração. Obrigado por cada sacrifício silencioso, por cada noite mal dormida ao meu lado, mesmo sem entender completamente a complexidade dos meus estudos, e por me ensinarem, com o exemplo, o verdadeiro significado do amor e da resiliência. Esta conquista é tão minha quanto de vocês.

Ao meu amor, Victor Cordeiro Simão, meu companheiro de jornada. Obrigado por segurar a minha mão nos dias tempestuosos, por enxugar minhas lágrimas de frustração e por celebrar com entusiasmo cada etapa vencida. Sua paciência infinita, seu colo acolhedor e seu amor incondicional foram o combustível que me manteve em movimento. Ter você ao meu lado tornou essa caminhada mais leve e feliz.

À minha orientadora, Prof.^a Dra. Graziela Torres Blanch, minha mais profunda admiração e gratidão. Obrigado por me guiar não apenas pelos caminhos da pesquisa científica, mas por me ensinar a ter olhar crítico, ética e paixão pelo que se faz. Sua paciência, seus conselhos precisos e sua dedicação foram fundamentais para a construção deste trabalho e para o meu crescimento como futura médica.

A todos os professores e profissionais que cruzaram meu caminho durante a graduação, deixando um pouco de si e contribuindo para a minha formação. Levarei comigo cada ensinamento.

Por fim, agradeço a cada paciente que cruzou meu caminho durante o curso. Vocês me ensinaram que a medicina vai muito além dos livros; ela se aprende na escuta ativa, no toque humanizado e no olhar que enxerga a pessoa por trás da doença.

Muito obrigada a todos que, de alguma forma, fizeram parte desta história.

RESUMO

INTRODUÇÃO: O Transtorno do Espectro Autista (TEA) é uma condição do neurodesenvolvimento de origem multifatorial, cuja prevalência tem aumentado globalmente, embora haja escassez de dados no Brasil sobre fatores de risco perinatais e particularidades regionais. **OBJETIVO:** Identificar fatores de risco pré e perinatais associados ao diagnóstico de TEA em crianças com até cinco anos no estado de Goiás, Brasil. **MÉTODOS:** Trata-se de um estudo caso-controle com linkage determinístico de bases de dados em saúde, utilizando o Sistema de Informações sobre Nascidos Vivos (SINASC) e o Registro das Ações Ambulatoriais de Saúde – Atenção Psicossocial (RAAS-PSI). A população incluiu nascidos vivos em Goiás no ano de 2019, com seguimento até 2024 para identificação de diagnósticos de TEA (CID-10 F84). Os controles foram selecionados por amostragem aleatória simples entre os nascidos sem o desfecho. Foram aplicadas técnicas de imputação múltipla por equações encadeadas (MICE) para dados ausentes e regressão logística hierárquica com seleção de variáveis por LASSO. **RESULTADOS:** Foram identificados 96.112 nascimentos e 1.033 casos de TEA, resultando em prevalência acumulada de 1,07% (IC95% 1,01-1,14). Na análise multivariada com seleção LASSO, o sexo masculino (OR=3,35; IC95% 2,86-3,92; $p<0,001$) e a presença de malformações congênitas (OR=2,48; IC95% 1,26-4,88; $p=0,008$) permaneceram significativamente associados ao diagnóstico de TEA. Variáveis clássicas da literatura como idade parental avançada, prematuridade e baixo peso ao nascer não apresentaram significância estatística nos modelos ajustados. **CONCLUSÃO:** O sexo masculino e as malformações congênitas constituem os principais fatores de risco para o diagnóstico precoce de TEA na população estudada. O linkage de bases de dados públicas mostrou-se ferramenta viável para produção de evidências nacionais, embora a homogeneidade socioeconômica da amostra possa ter mascarado outras associações. Os achados reforçam a necessidade de vigilância do neurodesenvolvimento em grupos de maior vulnerabilidade e de estratégias de diagnóstico sensíveis às particularidades de gênero.

Palavras-chave: Transtorno do Espectro Autista; Fatores de Risco; Linkage de Dados.

ABSTRACT

INTRODUCTION: Autism Spectrum Disorder (ASD) is a neurodevelopmental condition with multifactorial etiology, whose prevalence has increased globally, although data on perinatal risk factors and regional particularities remain scarce in Brazil. **OBJECTIVE:** To identify pre- and perinatal risk factors associated with ASD diagnosis in children up to five years of age in the state of Goiás, Brazil. **METHODS:** This was a case-control study using deterministic linkage of health databases, utilizing the Live Birth Information System (SINASC) and the Outpatient Health Actions Registry – Psychosocial Care (RAAS-PSI). The population included live births in Goiás during 2019, with follow-up until 2024 to identify ASD diagnoses (ICD-10 F84). Controls were selected through simple random sampling from births without the outcome. Multiple imputation by chained equations (MICE) was applied for missing data, and hierarchical logistic regression with LASSO variable selection was performed. **RESULTS:** A total of 96,112 births and 1,033 ASD cases were identified, yielding a cumulative prevalence of 1.07% (95%CI 1.01-1.14). In the multivariate analysis with LASSO selection, male sex (OR=3.35; 95%CI 2.86-3.92; $p<0.001$) and the presence of congenital malformations (OR=2.48; 95%CI 1.26-4.88; $p=0.008$) remained significantly associated with ASD diagnosis. Classic variables from the literature such as advanced parental age, prematurity, and low birth weight showed no statistical significance in the adjusted models. **CONCLUSION:** Male sex and congenital malformations constitute the main risk factors for early ASD diagnosis in the studied population. Public database linkage proved to be a feasible tool for producing national evidence, although the socioeconomic homogeneity of the sample may have masked other associations. The findings reinforce the need for neurodevelopmental surveillance in higher vulnerability groups and for diagnostic strategies sensitive to gender particularities.

Keywords: Autism Spectrum Disorder; Risk Factors; Data Linkage.

ÍNDICE DE ILUSTRAÇÕES

FIGURAS

Figura 1. Fluxograma do processamento e linkage determinístico dos dados.	20
---	----

TABELAS

Tabela 1. Número de casos e controles estudados, segmentado por fatores maternos, fatores gestacionais e fatores perinatais.	21
Tabela 2. Valores da odds ratio bruta (OR), do intervalo de confiança (IC 95%) e da significância (p) obtidos pela análise univariada por regressão logística para associação entre diagnóstico de TEA em crianças com 5 anos ou menos (nascidas em 2019) e os fatores de risco para o estado de Goiás.	24
Tabela 3. Valores da odds ratio ajustado (OR), do intervalo de confiança (IC 95%) e da significância (p) obtidos pela análise multivariada por regressão logística para associação entre diagnóstico de TEA em crianças com 5 anos ou menos (nascidas em 2019) e os fatores de risco para o estado de Goiás.	26
Tabela 4. Valores da <i>odds ratio</i> ajustado (OR), do intervalo de confiança (IC 95%) e da significância (p) obtidos pela análise multivariada com seleção das variáveis por LASSO (seleção por λ_{1se}) para associação entre diagnóstico de TEA em crianças com 5 anos ou menos (nascidas em 2019) e os fatores de risco para o estado de Goiás.	28

QUADROS

Quadro 1. Variáveis relacionadas às características maternas.	12
Quadro 2. Variáveis relacionadas às características do recém-nascido.	13

LISTA DE ABREVIATURAS E SIGLAS

APGAR	Escala de avaliação da vitalidade do recém-nascido
CAPSi	Centro de Atenção Psicossocial Infantojuvenil
CDC	<i>Centers for Disease Control and Prevention</i>
CEP	Comitê de Ética em Pesquisa
CNS	Conselho Nacional de Saúde
DSM-V	<i>Diagnostic and Statistical Manual of Mental Disorders</i>
GLM	<i>Generalized Linear Models</i>
LASSO	<i>Least Absolute Shrinkage and Selection Operator</i>
LRT	<i>Likelihood Ratio Test</i>
MAR	<i>Missing at Random</i>
MCAR	<i>Missing Completely at Random</i>
MICE	<i>Multivariate Imputation by Chained Equations</i>
OMS	Organização Mundial da Saúde
OR	<i>Odds Ratio</i>
PUC Goiás	Pontifícia Universidade Católica de Goiás
RAAS-PSI	Registros das Ações Ambulatoriais de Saúde – Atenção Psicossocial
RECORD	<i>REporting of studies Conducted using Observational Routinely-collected health Data</i>
RN	Recém-nascido
RWD	<i>Real World Data</i>
RWE	<i>Real World Evidence</i>
SAI	Sistema de Informações Ambulatoriais
SINASC	Sistema de Informações sobre Nascidos Vivos
STROBE	<i>STrengthening the Reporting of OBservational studies in Epidemiology</i>
SUS	Sistema Único de Saúde
TEA	Transtorno do Espectro Autista

SUMÁRIO

1. INTRODUÇÃO	8
2. OBJETIVOS	9
2.1. Objetivos secundários.....	9
3. METODOLOGIA.....	10
3.1. Tipo de estudo e constituição da amostra	10
3.1.1. Cenário	10
3.1.2. População.....	10
3.2. Variáveis	11
3.3. Fontes de dados	13
3.4. Linkage	13
3.5. Viés.....	14
3.6. Tamanho de amostra.....	14
3.7. Análise estatística	15
3.8. Considerações éticas.....	17
4. RESULTADOS	18
4.1. Linkage	18
4.2. Características da amostra	19
4.3. Testes estatísticos	22
5. DISCUSSÃO	29
6. CONCLUSÃO.....	35
REFERÊNCIAS	36
ANEXOS – LINHAS DE CÓDIGO EM R UTILIZADAS	43
1. Linkage	43
2. Formatação do banco de dados	46
3. MICE e Análise Univariada/Multivariada.....	49
4. LASSO.....	54

1. INTRODUÇÃO

O Transtorno do Espectro Autista (TEA) é uma condição definida por atraso de neurodesenvolvimento marcada por prejuízos biopsicossociais (Kim et al., 2019). Atualmente, entende-se que o quadro possui origem multifatorial, isto é, é determinado por aspectos genéticos e ambientais (Andrade, 2025; Qiu et al., 2022). O TEA é uma das principais causas de incapacidade em menores de 5 anos, caracteriza-se pelo surgimento de sintomas no período de 12 a 18 meses de idade, embora o diagnóstico clínico seja geralmente confirmado por volta dos dois anos (Baxter et al., 2015).

O TEA é caracterizado por déficits persistentes na comunicação social, pela presença de comportamentos, de atividades e de interesses repetitivos e restritos (Hodges; Fealko; Soares, 2020). Historicamente, homens possuem o diagnóstico mais frequentemente em relação às mulheres de 3 para 1, no entanto, este achado pode ser devido ao viés de gênero do diagnóstico (Loomes; Hull; Mandy, 2017a). Este subdiagnóstico em mulheres pode ser explicado por diferenças no neurodesenvolvimento entre os sexos ou por diferenças clínicas, como o “mascaramento” dos sintomas, que é mais prevalente em mulheres (Alaghband-Rad; Hajikarim-Hamedani; Motamed, 2023; Lockwood Estrin et al., 2021). Devido ao TEA ser um espectro, o DSM-V estabelece três níveis de gravidade: Nível 1, na ausência de apoio, os déficits na comunicação social causam prejuízos notáveis; Nível 2 prejuízos sociais são notáveis mesmo com apoio; e Nível 3, prejuízos graves nas competências de comunicação social verbal e não verbal causam grandes limitações e respostas mínima à estímulos sociais (American Psychiatric Association, 2022; Waizbard-Bartov et al., 2023).

Atualmente, dentre os aspectos ambientais, existem evidências de que fatores de risco perinatais, como idade parental avançada, hipóxia, prematuridade e parto cesárea podem estar relacionados ao desenvolvimento do TEA. Contudo, essas evidências são limitadas e com associações estatísticas que não confirmam relações causais, além da heterogeneidade metodológica entre os estudos, bem como divergências entre os resultados (Abdelkader *et al.*, 2024; Gardener; Spiegelman; Buka, 2009; Kim *et al.*, 2019; Sousamli *et al.*, 2024; Wang, C. *et al.*, 2017; Yang *et al.*, 2025).

Do ponto de vista epidemiológico, há um crescente aumento da prevalência do TEA globalmente (Zeidan et al., 2022). Estimativas indicam que uma em cada 36 crianças estadunidenses recebem o diagnóstico (Maenner et al., 2023), enquanto dados do Reino Unido sinalizam um aumento expressivo das taxas de incidência no diagnóstico de TEA nas últimas duas décadas (Russell et al., 2022). Apesar disso, alguns estudos ressaltam a escassez de dados em países de renda baixa e média. No Brasil, essa escassez é observada pelas poucas e limitadas evidências sobre a prevalência do TEA (Dellazari et al., 2024a; Paula et al., 2011). Esse contexto revela a necessidade urgente de pesquisas e intervenções adequadas para a demografia e particularidades regionais do Brasil relacionadas ao TEA.

Assim, o presente estudo buscou identificar fatores de risco pré e perinatais para o diagnóstico de TEA em crianças com 5 anos ou menos em uma Unidade Federativa do Brasil, a fim de apoiar o planejamento e desenvolvimento de recursos baseado em evidências nacionais e regionais para contribuir para políticas de saúde pública para intervenção e diagnóstico precoce de TEA.

2. OBJETIVOS

Identificar fatores de risco pré e perinatais associados ao diagnóstico de TEA em crianças com 5 anos ou menos em uma Unidade Federativa do Brasil.

2.1. Objetivos secundários

Identificar a prevalência de TEA em crianças com 5 anos ou menos em uma Unidade Federativa do Brasil, bem como sua correlação com fatores sociodemográficos.

3. METODOLOGIA

3.1. Tipo de estudo e constituição da amostra

Trata-se de um estudo do tipo caso-controle utilizando bases de dados em saúde, descrito conforme as orientações da ferramenta STROBE (*STrengthening the Reporting of OBservational studies in Epidemiology*). Devido à natureza do estudo, que se baseia em fontes de dados de saúde coletados rotineiramente para fins administrativos, foram também incorporadas as diretrizes da extensão RECORD (*REporting of studies Conducted using Observational Routinely-collected health Data*). A população do estudo foi estabelecida a partir dos registros históricos do ambiente do Sistema de Informações sobre Nascidos Vivos (SINASC/SUS) e o seguimento para identificação dos casos de interesse foi realizado por meio da análise dos Registros das Ações Ambulatoriais de Saúde – Atenção Psicossocial (RAAS-PSI), ocorridos nos anos subsequentes.

3.1.1. Cenário

O estudo foi realizado no contexto de uma única Unidade Federativa do Brasil, Goiás. Em 2019, Goiás apresentava uma cobertura de 98,5% no SINASC, além de uma rede de CAPSi estruturada, o que possibilita a análise dos dados.

3.1.2. População

A população foi composta pelos registros do SINASC. Foram considerados como casos os registros pareados entre o SINASC e o RAAS-PSI, contendo o desfecho de interesse. Foram incluídos todos nascidos vivos de mães residentes em Goiás, cujos nascimentos ocorreram no período de 1º de janeiro de 2019 a 31 de dezembro de 2019. Foram excluídos os registros com inconsistências graves ou ausência de dados em variáveis chaves para o processo de vinculação, como data de nascimento e sexo. Crianças menores de 1 ano foram excluídas devido à incerteza do diagnóstico ou devido às limitações no registro suspeito do diagnóstico nos sistemas de informação. Os controles foram obtidos por meio de amostra aleatória simples dos nascidos vivos que não tiveram o desfecho de interesse e, portanto, tendo como critério de inclusão não serem pareados com o banco de dados do RAAS-PSI. A função `slice_sample`, do pacote `Dplyr`, foi utilizada para gerar a amostra simples aleatória

3.2. Variáveis

Todas as variáveis de exposição e os potenciais fatores de confusão foram extraídos exclusivamente do banco de dados do SINASC. A seleção das variáveis foi guiada pelas evidências na literatura científica e a disponibilidade das variáveis no dicionário de dados oficial do SINASC. As variáveis foram divididas em 2 grupos: características maternas e características do recém-nascido. As variáveis categóricas tratadas foram nomeadas de acordo com seu nome no banco de dados original seguida do sufixo “_cat”, por exemplo “IDADEMAE_cat”

Quadro 1. Variáveis relacionadas às características maternas.

Grupo	Variável²	Descrição	Tratamento	Fator de Risco
Fatores maternos/histórico	ESMAE	Escolaridade da mãe	Categórica (3 anos ou menos, 4 a 7 anos, 8 a 11 anos de estudo ¹)	Baixa escolaridade materna
	ESTCIVMAE	Situação conjugal da mãe	Categórica (Com parceiro ¹ , Sem parceiro)	Suporte social
	RACACORMAE	Raça/cor da mãe	Categórica (Branca ¹ , Preta/Parda)	Fator sociodemográfico
	IDADEMAE	Idade da mãe em anos	Categórica (< 30 anos, 30 e 34 anos ¹ , ≥ 35 anos)	Idade materna avançada
	IDADEPAI	Idade do Pai	Contínua e Categórica (< 30 anos, 30 e 34 anos ¹ , 35 e 40 anos, ≥ 40 anos)	Idade paterna avançada
Fatores gestacionais	QTDFILVIVO	Nº de filhos vivos	Categórica (Nenhum ¹ , 1 ou mais)	Paridade
	QTDFILMORT	Nº de filhos mortos	Categórica (Nenhum ¹ , 1 ou mais)	Histórico de perdas gestacionais
	CONSULTAS	Nº de consultas pré-natal	Categórica (0-6 - Inadequado, ≥ 7 - Adequado ¹)	Cuidado pré-natal inadequado
	GESTACAO	Semanas de gestação	Categórica (<37 semanas: pré-termo; ≥ 37: termo ¹)	Prematuridade
	TPGRAVIDEZ	Tipo de gravidez	Categórica (Única ¹ , Múltipla)	Gemelaridade
	PARTO	Tipo de parto	Categórica (Vaginal ¹ , Operatório)	Parto cesáreo

Fonte: Elaboração própria, 2025. ¹Categoria de Referência. ²Nome da variável no banco de dados antes da categorização.

Quadro 2. Variáveis relacionadas às características do recém-nascido.

Grupo	Variável ²	Descrição	Tratamento	Fator de Risco
Fatores perinatais	APGAR1	Escore de Apgar no 1º minuto	Categórica (≤ 7 e $> 7^1$)	Sofrimento fetal/Hipóxia perinatal
	APGAR5	Escore de Apgar no 5º minuto	Categórica (≤ 7 , $> 7^1$)	Sofrimento fetal/Hipóxia perinatal
	SEXO	Sexo do RN	Categórica (Masculino, Feminino ¹)	Sexo masculino
	PESO	Peso ao nascer em gramas	Categórica ($<2500g$ e $\geq 2500g^1$)	Baixo peso ao nascer
	IDANOMAL	Anomalia congênita detectada	Categórica (Sim, Não ¹)	Malformações congênitas

Fonte: Elaboração própria, 2025. ¹Categoria de referência. ²Nome da variável no banco de dados antes da categorização.

3.3. Fontes de dados

O estudo utilizou exclusivamente dados secundários provenientes do SINASC e do ambiente do Sistema de Informações Ambulatoriais do SUS (SIA/SUS), em seu subsistema RAAS-PSI, com registro dos atendimentos nos CAPSi. Os registros de diagnóstico no RAAS-PSI, em sua variável CIDPRI, utilizam a codificação segundo a Classificação Estatística Internacional de Doenças e Problemas Relacionados com Saúde, 10ª Revisão (CID-10). O desfecho de interesse foi considerado como a ocorrência do primeiro registro de atendimento ambulatorial em um CAPSi, identificado na base de dados do subsistema RAAS-PSI, que contenha um dos seguintes códigos da CID-10 no campo CIDPRI: CID-10 F84, F84.0, F84.1, F84.5, F84.8 e F84.9. O desfecho é um registro administrativo de diagnóstico, o que implica em um potencial viés de classificação.

3.4. Linkage

O relacionamento dos registros individuais entre as bases de dados do SINASC e do SIA/SUS (em seu subsistema RAAS-PSI) foi realizado com uma estratégia de linkage determinístico. O processamento dos dados foi executado com uso de softwares RStudio® e seu pacote *Record Linkage Toolkit* (Fellegi; Sunter, 1969; Van Der Laan, 2022).

Para otimizar o processo de comparação e reduzir a carga computacional, foi aplicada a estratégia de bloqueio (*blocking*), na qual apenas os registros que coincidirem em chaves comuns foram comparados. A data de nascimento (DTNASC) foi a chave de bloqueio selecionada para este estudo. As variáveis de comparação foram a data de nascimento (DTNASC), o sexo (SEXO) e a raça/cor (RACACOR).

Previamente ao linkage, as variáveis de identificação selecionadas foram submetidas a uma padronização. Este processo incluiu a formatação de datas para o padrão ISO 8601 e a harmonização dos valores das colunas de sexo entre as bases (ex: transformação de "1" e "2" do SINASC para "M" e "F", e renomeação da coluna "SEXOPAC" do RAAS-PSI para "SEXO").

Um par de registros (um do SINASC e um do RAAS-PSI) foi classificado como um "par casado" (um match) somente quando houve concordância exata em todas as três chaves de identificação (DTNASC, SEXO e RACACOR). Os registros do SINASC que não encontraram correspondência no RAAS-PSI constituíram a população base para a amostragem dos controles.

A variável "CONTADOR" é única e permitiu a identificação e remoção de duplicatas. O algoritmo de comparação utilizado para as variáveis de DTNASC e SEXO foi de concordância exata. A acurácia foi avaliada pelas funções do pacote *Record Linkage Toolkit*.

3.5. Viés

Foram considerados e abordados os seguintes vieses: (a) viés de classificação errada do desfecho; (b) viés de seleção por falha do *linkage*; (c) viés de detecção por acesso desigual aos serviços.

3.6. Tamanho de amostra

O tamanho amostral foi calculado para ser possível identificar significância de odds ratio (OR) maior que 1.5, considerando um poder de estudo ($1-\beta$) de 80%, um erro α de 5% e uma frequência relativa de 8.7% entre os controles, adotando-se como fator o baixo peso ao nascer ($< 2500\text{g}$). Optou-se pela seleção de quatro controles para cada caso ($k = 4$). Portanto, foram necessários 664 casos e 2655 controles.

3.7. Análise estatística

Inicialmente, foi realizada uma análise descritiva dos bancos de dados gerados. As características dos casos e controles foram sumarizadas utilizando frequências absolutas e relativas para variáveis categóricas. Para variáveis contínuas foram utilizadas medidas de tendência central (média) e de dispersão (desvio padrão). A prevalência (incidência cumulativa) de diagnóstico de TEA aos 5 anos de idade ou menos foi calculada como a proporção de crianças que tiveram desfecho até essa idade, com seu respectivo intervalo de confiança de 95% (IC 95%).

As variáveis foram analisadas de forma univariada e aquelas com $p < 0,25$ foram incluídas na análise multivariada por regressão logística binária, utilizando-se o pacote GLM no RStudio®. Na análise univariada, foi verificada a associação entre o desfecho e cada variável independente, com cálculo de significância obtido pela estatística do Qui-quadrado de Pearson. O modelo multivariado de regressão logística elaborado no RStudio® forneceu os valores finais da razão de chances, denominada odds ratio (OR) com intervalo de confiança de 95%. O modelo foi hierarquizado, de tal forma que contemplou as interações entre os diversos grupos de fatores. Portanto, a regressão foi feita em cinco etapas: as três primeiras foram às regressões multivariadas dentro de cada bloco (fatores maternos/históricos, fatores gestacionais e fatores perinatais). Em cada etapa, foram incluídas as variáveis com $p < 0,25$ na análise univariada. A permanência no modelo final de cada bloco foi determinada pela relevância teórica da variável e pelo $p < 0,10$. Na quarta etapa, foram incorporadas as variáveis do modelo final de fatores maternos/históricos àquelas do modelo final de fatores gestacionais, com a permanência das variáveis com relevância teórica ou $p < 0,10$. Na última etapa, as variáveis do modelo anterior foram unidas as variáveis de fatores perinatais, mantendo as variáveis com relevância teórica ou $p < 0,10$.

Além disso, para seleção das variáveis a serem incluídas nos modelos de regressão logística hierárquica, foi realizada uma abordagem de regularização, em complementação ao critério de seleção univariado ($p < 0,25$). Foi empregada a técnica LASSO (*Least Absolute Shrinkage and Selection Operator*), um método de regressão penalizada que se destaca por sua capacidade de selecionar um subconjunto parcimonioso de variáveis, sendo robusto à multicolinearidade. O LASSO funciona aplicando uma penalidade que "encolhe" os coeficientes (Odds Ratios) de todas as

variáveis, forçando os preditores menos relevantes a exatamente zero, efetivamente removendo-os do modelo. Foi utilizado o primeiro banco de dados completo gerado pelo processo de MICE. Todas as variáveis categóricas (ex: SEXO_cat, PESO_cat) foram transformadas em variáveis *dummy* (0/1) para compor a matriz de preditores. Foi ajustado um modelo LASSO ($\alpha=1$) para regressão logística binomial (*family="binomial"*) utilizando validação cruzada (*k-fold cross-validation*) para identificar o nível ideal de penalização (*lambda*). Foi adotado o critério *lambda.1se*. As variáveis que mantiveram coeficientes diferentes de zero no modelo *lambda.1se* foram selecionadas. As variáveis identificadas por este processo foram utilizadas na análise de regressão logística hierárquica subsequente.

A proporção e o padrão de dados ausentes foram avaliados para cada variável. Foi empregado a técnica de imputação múltipla por equações encadeadas (MICE), por meio do pacote *mice*. A técnica foi utilizada em casos de dados ausentes em covariáveis relevante para o modelo multivariado que excederam em 5% e com evidências de que a ausência é completamente aleatória (*Missing Completely at Random – MCAR*) ou aleatória (*Missing at Random – MAR*).

Para o *mice*, foi configurada a criação de 5 bancos de dados completos com os dados imputados nas variáveis processadas. Nesses casos, a análise univariada foi realizada com o Teste de Razão de Verossimilhança (LRT), que consiste no ajuste e combinação de dois modelos (nulo e univariado) de regressão logística nos bancos de dados imputados pelo MICE.

Todos os testes realizados foram feitos ao nível de significância $\alpha = 0,05$.

3.8. Considerações éticas

Não houve necessidade de submissão do projeto do estudo ao Comitê de Ética em Pesquisa da Pontifícia Universidade Católica de Goiás (PUC Goiás), uma vez que se trata de pesquisa sobre dados secundários, sem identificação dos participantes, conforme prevê a Resolução do Conselho Nacional de Saúde (CNS) no 466, de 12 de dezembro de 2012, e a Resolução CNS no 510, de 7 de abril de 2016 – esta segunda resolução dispõe sobre a não necessidade de submissão a um CEP em pesquisas que utilizem informações de domínio público.

4. RESULTADOS

Foram identificados 96.112 nascimentos no estado de Goiás entre 1 de janeiro de 2019 e 31 de dezembro de 2019. Quanto aos atendimentos, foram registrados 1.048.575 atendimentos no mesmo período. Os registros do RAAS-PSI foram filtrados inicialmente para nascidos no ano de 2019, totalizando 3.626 registros. Após o processamento inicial, os bancos de dados obtidos foram relacionados por meio do *linkage* determinístico, com a variável de bloqueio DTNASC e as variáveis de comparação SEXO+RACACOR. Obteve-se 3.626 registros pareados, ou seja 100% dos nascidos em 2019 que tiveram atendimentos no CAPSi (entre 2019 e 2024) foram pareados. Aplicou-se um filtro para atendimentos em crianças maiores de 1 ano, obtendo-se como resultado final 3.504 registros únicos. As entradas não pareadas (92.486) foram utilizadas para o banco de dados de controles (Figura 1).

4.1. Linkage

Como produto final das etapas de *linkage*, obteve-se o grupo de casos, constituído por 1.033 diagnósticos com um dos seguintes códigos da CID-10 na variável CIDPRI: F84, F84.0, F84.1, F84.5, F84.8 e F84.9. O grupo de controles foi composto por uma amostra de 4.130 crianças não pareadas com o RAAS-PSI. Os valores foram superiores aos mínimos estipulados para o tamanho da amostra (Figura 1).

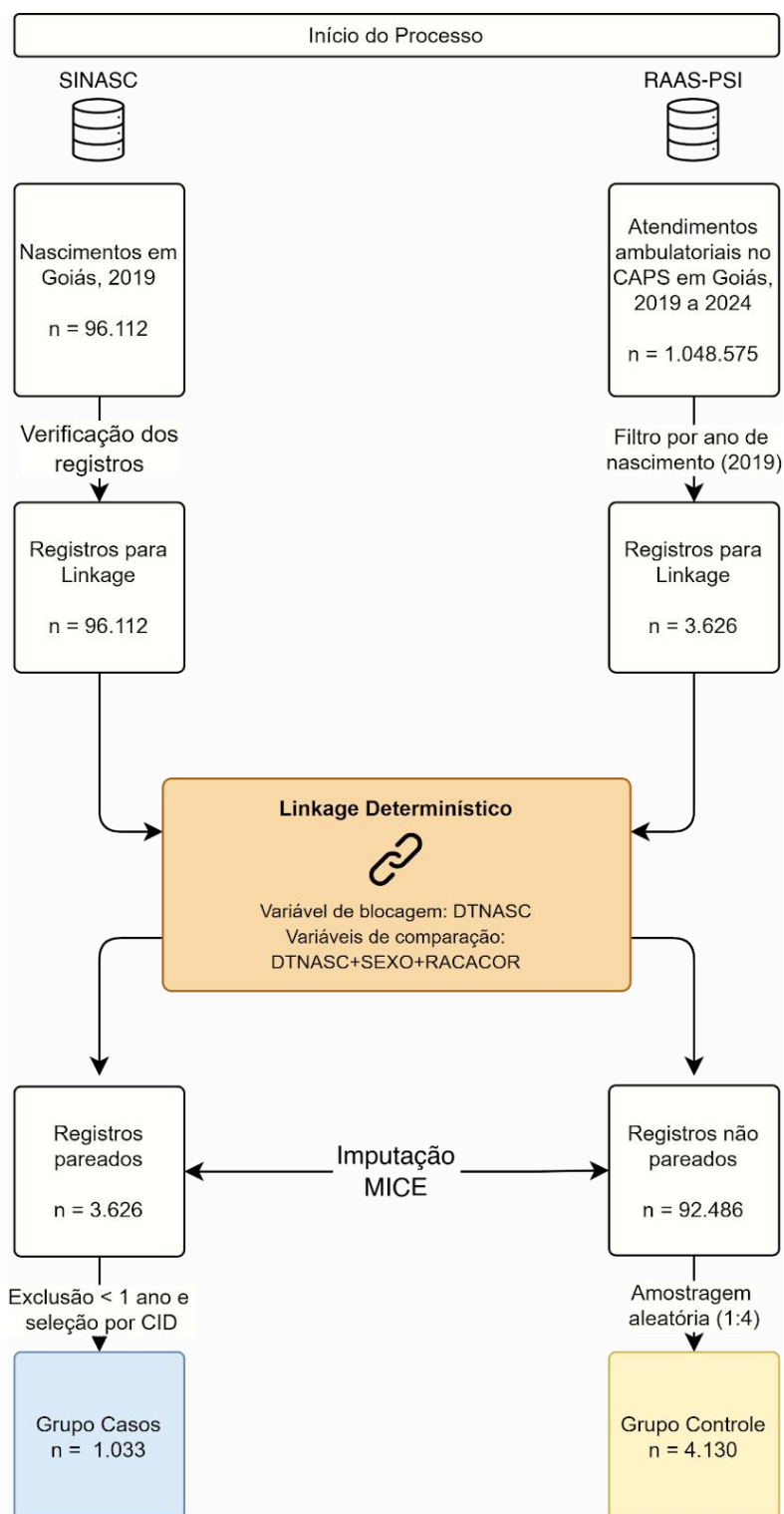
Apenas a variável idade paterna (IDADEPAI) estava com mais de 5% dos dados ausentes. Contabilizamos 3.111 (60.2%) registros vazios de idade paterna, sendo 2.569 (62%) nos controles e 542 (52%) nos casos. Dessa forma, primeiro foi realizada uma avaliação do padrão de ausência ($p < 0.001$) que indicou que os dados eram ausentes ao acaso (MAR), o que significa que a probabilidade de um valor estar faltando depende apenas dos valores observados. Visto que a proporção de ausência excedia 5% e se enquadra na premissa de MAR, foi empregada a técnica MICE para preencher os valores faltantes. Este processo gerou cinco bancos de dados imputados que serviram de base para as análises subsequentes.

As variáveis de interesse foram categorizadas para obtenção das categorias necessárias para avaliação, conforme mencionado no Quadro 1 e 2. A categorização foi feita com auxílio do pacote *Dplyr* no RStudio®.

Como a variável idade paterna passou por imputações múltiplas, sua categorização (IDADEPAI_cat), não foi submetida diretamente ao Teste Qui-quadrado de Pearson, pois este não considera a incerteza dos dados imputados. Em substituição, foi utilizado um Teste de Razão de Verossimilhança (LRT) combinado. Esta abordagem consistiu em ajustar e combinar os resultados de dois modelos de regressão logística em todos os bancos de dados imputados: um modelo nulo (apenas intercepto) e um modelo univariado (contendo a variável categórica como único preditor). O p-valor obtido da comparação estatística entre esses dois modelos combinados ("poolados") representa a significância da associação univariada, incorporando adequadamente a incerteza introduzida pelo processo de imputação. O resultado foi de $p=0.533$.

4.2. Características da amostra

Em nossos dados, quanto às variáveis relacionadas aos fatores maternos, identificamos uma predominância de mães com escolaridade de 8 a 11 anos (88%), com parceiro (53%), parda/preta/amarela (72%), com idade menor do que 30 anos (65%). A maior parte dos parceiros apresentavam idade também menor do que 30 anos (43%). A maioria das mães apresentavam um ou mais filhos nascidos vivos (59%) e a maioria também apresentava nenhum filho nascido morto (80%). Em relação aos fatores gestacionais, o acompanhamento pré-natal foi adequado (com quantidade maior do que 6 consultas) em 72% dos registros. A gestação foi a termo (≥ 37 semanas) em 89% dos registros, com a maioria (98%) de gestações únicas, com parto operatório em 68% dos registros. Identificamos nos fatores perinatais que a maioria dos recém-nascidos apresentaram APGAR de 1° (87%) e 5° minuto (98%) maior do que 7. Nos registros, o sexo masculino foi predominante (56%). A maioria dos recém-nascidos não apresentaram baixo peso (92%). Quase a totalidade dos registros (99%) não apresentaram malformações congênitas (Tabela 1). A prevalência (ou incidência acumulada) de diagnóstico de Transtorno do Espectro Autista (TEA) em crianças de até 5 anos no estado de Goiás foi de 1.07% (IC 95% 1.01-1.14)

Figura 1. Fluxograma do processamento e linkage determinístico dos dados.

Fonte: Elaboração própria, 2025.

Tabela 1. Número de casos e controles estudados, segmentado por fatores maternos, fatores gestacionais e fatores perinatais.

Variáveis	Total, n=5163 ¹	Controles, n=4130 ¹	Casos, n=1033 ¹	p ²
<i>Fatores maternos/histórico</i>				
Escolaridade materna				0.423
8 a 11 anos	4512 / 5130 (88%)	3601 / 4101 (88%)	911 / 1029 (89%)	
4 a 7 anos	571 / 5130 (11%)	465 / 4101 (11%)	106 / 1029 (10%)	
3 anos ou menos	47 / 5130 (0.9%)	35 / 4101 (0.9%)	12 / 1029 (1.2%)	
Dados ausentes	33	29	4	
Estado civil da mãe				0.092
Com parceiro	2713 / 5111 (53%)	2147 / 4090 (52%)	566 / 1021 (55%)	
Sem parceiro	2398 / 5111 (47%)	1943 / 4090 (48%)	455 / 1021 (45%)	
Dados ausentes	52	40	12	
Raça/Cor				0.736
Branca	1360 / 4883 (28%)	1091 / 3902 (28%)	269 / 981 (27%)	
Parda/Preta	3523 / 4883 (72%)	2811 / 3902 (72%)	712 / 981 (73%)	
Dados ausentes	280	228	52	
Idade da mãe (anos)				0.822
Entre 30 e 34 anos	1056 / 5163 (20%)	842 / 4130 (20%)	214 / 1033 (21%)	
< 30 anos	3351 / 5163 (65%)	2677 / 4130 (65%)	674 / 1033 (65%)	
≥ 35 anos	756 / 5163 (15%)	611 / 4130 (15%)	145 / 1033 (14%)	
Idade do pai (anos)				0.533 ³
Entre 30 e 34 anos	548 / 2167 (25%)	434 / 1676 (26%)	114 / 491 (23%)	
< 30 anos	926 / 2167 (43%)	710 / 1676 (42%)	216 / 491 (44%)	
Entre 35 e 40 anos	382 / 2167 (18%)	298 / 1676 (18%)	84 / 491 (17%)	
≥ 40 anos	311 / 2167 (14%)	234 / 1676 (14%)	77 / 491 (16%)	
Dados ausentes	2996	2454	542	
Paridade (filhos nascidos vivos)				0.702
Nenhum	2113 / 5133 (41%)	1694 / 4102 (41%)	419 / 1031 (41%)	
1 ou mais	3020 / 5133 (59%)	2408 / 4102 (59%)	612 / 1031 (59%)	
Dados ausentes	30	28	2	
Paridade (filhos nascidos mortos)				0.083
Nenhum	4070 / 5074 (80%)	3266 / 4047 (81%)	804 / 1027 (78%)	
1 ou mais	1004 / 5074 (20%)	781 / 4047 (19%)	223 / 1027 (22%)	
Dados ausentes	89	83	6	
<i>Fatores gestacionais</i>				
Número de consultas				0.185
Adequado	3667 / 5127 (72%)	2911 / 4094 (71%)	756 / 1033 (73%)	
Inadequado	1460 / 5127 (28%)	1183 / 4094 (29%)	277 / 1033 (27%)	
Dados ausentes	36	36	0	
Duração da gestação (semanas)				0.855
≥ 37 Semanas	4601 / 5149 (89%)	3679 / 4119 (89%)	922 / 1030 (90%)	
< 37 Semanas	548 / 5149 (11%)	440 / 4119 (11%)	108 / 1030 (10%)	
Dados ausentes	14	11	3	
Tipo de gravidez				0.521
Única	5058 / 5161 (98%)	4045 / 4130 (98%)	1013 / 1031 (98%)	

Múltipla	103 / 5161 (2.0%)	85 / 4130 (2.1%)	18 / 1031 (1.7%)	
Dados ausentes	2	0	2	
Tipo de parto				0.630
Vaginal	1672 / 5163 (32%)	1331 / 4130 (32%)	341 / 1033 (33%)	
Operatório	3491 / 5163 (68%)	2799 / 4130 (68%)	692 / 1033 (67%)	
<i>Fatores perinatais</i>				
APGAR 1º minuto				0.055
> 7	4462 / 5121 (87%)	3583 / 4091 (88%)	879 / 1030 (85%)	
≤ 7	659 / 5121 (13%)	508 / 4091 (12%)	151 / 1030 (15%)	
Dados ausentes	42	39	3	
APGAR 5º minuto				0.221
> 7	5003 / 5121 (98%)	4002 / 4091 (98%)	1001 / 1030 (97%)	
≤ 7	118 / 5121 (2.3%)	89 / 4091 (2.2%)	29 / 1030 (2.8%)	
Dados ausentes		42	39	
Sexo do RN				<0.001
Feminino	2297 / 5163 (44%)	2061 / 4130 (50%)	236 / 1033 (23%)	
Masculino	2866 / 5163 (56%)	2069 / 4130 (50%)	797 / 1033 (77%)	
Peso do RN				0.067
≥ 2500g	4730 / 5163 (92%)	3769 / 4130 (91%)	961 / 1033 (93%)	
< 2500g	433 / 5163 (8.4%)	361 / 4130 (8.7%)	72 / 1033 (7.0%)	
Malformação congênita				<0.001
Não	4875 / 4910 (99%)	3879 / 3899 (99%)	996 / 1011 (99%)	
Sim	35 / 4910 (0.7%)	20 / 3899 (0.5%)	15 / 1011 (1.5%)	
Dados ausentes	253	231	22	

¹n / N (%); ²Qui-quadrado de Pearson; ³Idade do pai (anos), com 52.3% de dados ausentes, foi utilizada Imputação Múltipla, p-valor calculado pelo Teste de Razão de Verossimilhança combinado sobre os dados imputados.

4.3. Testes estatísticos

Em relação à análise estatística, dentre os fatores maternos/histórico, a escolaridade materna de 4 a 7 anos (p=0.351) ou 3 anos ou menos (p=0.380), o estado civil (p= 0.097), a raça/cor parda/preta (p=0.708), a idade da mãe < 30 anos (p=0.915) ou ≥ 35 anos (p=0.568), a idade do pai < 30 anos (p=0.861), entre 35 e 40 anos (p=0.518) ou ≥ 40 anos (p=0.279), a paridade de filhos nascidos vivos (p=0.713) e a paridade de filhos nascidos mortos (p=0.099) não foram associadas ao aumento ou redução do risco de apresentar o diagnóstico de TEA em crianças até os 5 anos (Tabela 2).

Quanto aos fatores gestacionais, também não houve aumento ou redução do risco de apresentar diagnóstico de TEA em crianças até os 5 anos em relação às

variáveis número de consultas, duração da gestação, tipo de gravidez e tipo de parto (Tabela 2).

Dentre os fatores perinatais, os indivíduos do sexo masculino tiveram 3.364 vezes mais chances de apresentar o diagnóstico de TEA até os 5 anos (OR não ajustado: 3.364; IC 95% 2.874-3.938; $p < 0.001$). Da mesma forma, os indivíduos com malformação congênita tiveram quase 3 vezes mais chances de apresentar o diagnóstico de TEA até os 5 anos (OR não ajustado: 2.939; IC 95% 1.532-5.637; $p < 0.001$). As demais variáveis, APGAR 1º minuto, APGAR 5º minutos e peso do RN não foram associadas ao aumento ou diminuição das chances de apresentar o diagnóstico de TEA até os 5 anos (Tabela 2).

Tabela 2. Valores da odds ratio bruta (OR), do intervalo de confiança (IC 95%) e da significância (p) obtidos pela análise univariada por regressão logística para associação entre diagnóstico de TEA em crianças com 5 anos ou menos (nascidas em 2019) e os fatores de risco para o estado de Goiás.

Variáveis	OR ¹	IC 95%	p
<i>Fatores maternos/histórico</i>			
Escolaridade materna			
8 a 11 anos	1.000	-	-
4 a 7 anos	0.899	0.719-1.124	0.351
3 anos ou menos	1.344	0.695-2.599	0.380
Estado civil da mãe			
Com parceiro	1.000	-	-
Sem parceiro	0.889	0.774-1.022	0.097
Raça/Cor			
Branca	1.000	-	-
Parda/Preta	1.03	0.883-1.202	0.708
Idade da mãe (anos)			
Entre 30 e 34 anos	1.000	-	-
< 30 anos	0.991	0.834-1.177	0.915
≥ 35 anos	0.934	0.738-1.181	0.568
Idade do pai (anos)			
Entre 30 e 34 anos	1.000	-	-
< 30 anos	1.019	0.826-1.255	0.861
Entre 35 e 40 anos	1.089	0.836-1.418	0.518
≥ 40 anos	1.202	0.846-1.707	0.279
Paridade (filhos nascidos vivos)			
Nenhum	1.000	-	-
1 ou mais	1.026	0.893-1.179	0.713
Paridade (filhos nascidos mortos)			
Nenhum	1.000	-	-
1 ou mais	1.152	0.974-1.363	0.099
<i>Fatores gestacionais</i>			
Número de consultas			
Adequado	1.000	-	-
Inadequado	0.913	0.783-1.064	0.243
Duração da gestação (semanas)			
≥ 37 Semanas	1.000	-	-
< 37 Semanas	0.973	0.779-1.216	0.812
Tipo de gravidez			
Única	1.000	-	-
Múltipla	0.844	0.505-1.41	0.517
Tipo de parto			
Vaginal	1.000	-	-
Operatório	0.965	0.835-1.116	0.631
<i>Fatores perinatais</i>			
APGAR 1º minuto			

> 7	1.000	-	-
≤ 7	1.211	0.994-1.474	0.057
APGAR 5º minuto			
> 7	1.000	-	-
≤ 7	1.306	0.853-1.997	0.219
Sexo do RN			
Feminino	1.000	-	-
Masculino	3.364	2.874-3.938	< 0.001
Peso do RN			
≥ 2500g	1.000	-	-
< 2500g	0.782	0.601-1.017	0.067
Malformação congênita			
Não	1.000	-	-
Sim	2.939	1.532-5.637	< 0.001

¹OR não ajustado calculado por regressão logística univariada.

Na análise multivariada com seleção por blocos, as variáveis selecionadas no modelo final foram a escolaridade materna, a idade da mãe, a idade do pai, a duração da gestação (em semanas), o APGAR 1º minuto, o APGAR 5º minuto, o sexo do RN, o peso do RN e a presença de malformação congênita. Na análise ajustada, a única variável significativa foi o sexo masculino, que apresentou 2.506 vezes mais chances de receber o diagnóstico de TEA até os 5 anos (OR ajustado: 2.506; IC 95% 2.150-2.923; $p < 0.001$). As demais variáveis não foram associadas ao aumento ou redução das chances de apresentar o diagnóstico de TEA até os 5 anos (Tabela 3).

Tabela 3. Valores da odds ratio ajustado (OR), do intervalo de confiança (IC 95%) e da significância (p) obtidos pela análise multivariada por regressão logística para associação entre diagnóstico de TEA em crianças com 5 anos ou menos (nascidas em 2019) e os fatores de risco para o estado de Goiás.

Variáveis	OR ¹	IC 95%	p
<i>Fatores maternos/histórico</i>			
Escolaridade materna			
8 a 11 anos	1.000	-	-
4 a 7 anos	0.929	0.747-1.157	0.513
3 anos ou menos	0.939	0.050-0.175	0.842
Estado civil da mãe			
Com parceiro	1.000	-	-
Sem parceiro	-	-	-
Raça/Cor			
Branca	1.000	-	-
Parda/Preta	-	-	-
Idade da mãe (anos)			
< 20 anos	1.000	-	-
Entre 20 e 34 anos	1.092	0.889-1.341	0.400
≥ 35 anos	1.034	0.778-1.375	0.815
Idade do pai (anos)			
< 20 anos	1.000	-	-
Entre 20 e 34 anos	0.722	0.458-1.137	0.156
≥ 35 anos	0.773	0.462-1.292	0.313
Paridade (filhos nascidos vivos)			
Nenhum	1.000	-	-
1	-	-	-
≥ 2	-	-	-
Paridade (filhos nascidos mortos)			
Nenhum	1.000	-	-
1	-	-	-
≥ 2	-	-	-
<i>Fatores gestacionais</i>			
Número de consultas			
Adequado	1.000	-	-
Inadequado	-	-	-
Duração da gestação (semanas)			
≥ 37 Semanas	1.000	-	-
< 37 Semanas	1.014	0.788-1.305	0.916
Tipo de gravidez			
Única	1.000	-	-
Múltipla	-	-	-
Tipo de parto			
Vaginal	1.000	-	-
Operatório	-	-	-

*Fatores perinatais***APGAR 1º minuto**

> 7	1.000	-	-
≤ 7	1.181	0.958-1.454	0.120

APGAR 5º minuto

> 7	1.000	-	-
≤ 7	1.172	0.740-1.855	0.498

Sexo do RN

Feminino	1.000	-	-
Masculino	2.506	2.150-2.923	< 0.001

Peso do RN

≥ 2500g	1.000	-	-
< 2500g	0.784	0.580-1.060	0.113

Malformação congênita

Não	1.000	-	-
Sim	1.661	0.920-3.001	0.093

¹OR ajustado calculado por regressão logística multivariada.

Com o uso da regressão LASSO, com o critério lambda.1se, foram selecionadas as variáveis idade do pai, sexo do RN e a presença de malformação congênita. As demais variáveis foram reduzidas à zero pelo modelo com a penalização. Assim, a análise multivariada, com a seleção das variáveis por LASSO, indicou o sexo masculino associado a um risco de cerca de 3 vezes (OR ajustado: 3.348, IC 95%: 2.860-3.921, $p < 0.001$) em relação ao sexo feminino de apresentar o diagnóstico de TEA em crianças de 5 anos ou menos. A presença de malformação congênita também foi associada a um risco de apresentar o diagnóstico de TEA (OR ajustado: 2.482, IC 95%: 1.261-4.884, $p = 0.008$). A idade do pai não apresentou significância estatística, apesar de ter sido considerada uma variável relevante para ajuste, conforme a seleção por LASSO (Tabela 4).

Tabela 4. Valores da *odds ratio* ajustado (OR), do intervalo de confiança (IC 95%) e da significância (p) obtidos pela análise multivariada com seleção das variáveis por LASSO (seleção por λ_{1se}) para associação entre diagnóstico de TEA em crianças com 5 anos ou menos (nascidas em 2019) e os fatores de risco para o estado de Goiás.

Variáveis	OR ¹	IC 95%	p
<i>Fatores maternos/histórico</i>			
Idade do pai (anos)			
Entre 30 e 34 anos	1.000	-	-
< 30 anos	1.044	0.849-1.284	0.677
Entre 35 e 40 anos	1.093	0.846-1.413	0.488
≥ 40 anos	1.205	0.838-1.734	0.287
<i>Fatores perinatais</i>			
Sexo do RN			
Feminino	1.000	-	-
Masculino	3.348	2.860-3.921	<0.001
Malformação congênita			
Não	1.000	-	-
Sim	2.482	1.261-4.884	0.008

¹OR ajustado calculado por regressão logística multivariada com seleção das variáveis por LASSO.

O modelo de seleção por blocos divergiu da seleção LASSO nas variáveis escolaridade materna, idade da mãe, duração da gestação, APGAR 1º minuto, APGAR 5º minuto e peso do RN. A inclusão das variáveis no modelo final da seleção por blocos também considerou a relevância teórica das variáveis para ajuste de fatores de confusão, diferentemente da seleção LASSO que aplica um algoritmo para realizar esse ajuste.

5. DISCUSSÃO

Até onde sabemos, este é o primeiro estudo com uso da metodologia de *linkage* para identificar fatores de risco relacionados ao diagnóstico de TEA, com uso de informações de bancos de dados em saúde. Além disso, a abordagem obteve múltiplas variáveis de interesse como potenciais fatores de risco, identificadas em outros estudos. Ao realizar o *linkage* de bases de dados em saúde em uma coorte de nascidos vivos em Goiás no ano de 2019, identificamos o sexo masculino e a presença de malformações congênitas como os principais fatores associados ao diagnóstico de TEA em crianças menores de 5 anos. A utilização de métodos estatísticos robustos, como o MICE e o LASSO, permitiu mitigar a influência de problemas comuns em estudos com dados secundários, tais como a ausência de dados e a multicolinearidade. Diferentemente de parte da literatura, fatores de risco clássicos como idade parental avançada, prematuridade e baixo peso ao nascer não mantiveram significância estatística nos modelos finais ajustados nesta população específica. Ainda, o estudo gera Evidências de Mundo Real (RWE) sobre os fatores de risco para o TEA em uma região do Brasil, analisando Dados de Mundo Real (RWD) provenientes dos sistemas de informação em saúde do Brasil (SINASC e SIA/SUS).

A prevalência do diagnóstico de TEA em crianças de até 5 anos foi de aproximadamente 1,07% na região de coleta da amostra. O achado é próximo ao dados revelados pelo Censo Demográfico 2022 do Instituto Brasileiro de Geografia e Estatística (Instituto Brasileiro de Geografia e Estatística, 2025). Embora o número seja inferior às estimativas mais recentes do CDC estadunidense, que indicam uma prevalência de 1 em 31 crianças (3.2%), o crescimento expressivo nas últimas décadas pode sugerir um fortalecimento das políticas públicas de saúde mental e pela maior conscientização da sociedade e dos profissionais de saúde (Almeida; Neves, 2020; Dellazari *et al.*, 2024b; Dourado Souza Akahosi Fernandes *et al.*, 2025; Sukiennik; Marchezan; Scornavacca, 2022).

Em nosso estudo, na análise bivariada e multivariada (mesmo após ajuste LASSO) identificamos que RNs do sexo masculino tiveram entre 2.5 a 3.3 vezes mais chances de serem diagnosticados com TEA até os 5 anos. No mesmo sentido deste achado, revisões sistemáticas também indicaram uma maior proporção de

diagnósticos em crianças do sexo masculino (Christensen *et al.*, 2019; Loomes; Hull; Mandy, 2017b; Shaw *et al.*, 2025). No entanto, discute-se se o diagnóstico precoce, predominante do sexo masculino, está relacionado à fatores intrínsecos (como fatores biológicos e inerentes ao indivíduo) e extrínsecos (como fatores ambientais ou viés externo no processo diagnóstico), uma vez que, por exemplo, dificuldades de expressões na linguagem podem ser vistas mais precocemente nos meninos em relação às meninas (Dillon *et al.*, 2023; Halladay *et al.*, 2015; Lai; Baron-Cohen; Buxbaum, 2015; Mandy; Lai, 2017). Além disso, meninas com dificuldades de comunicação social podem ser subdiagnosticadas devido ao uso de estratégias de camuflagem, o que atrasa a identificação do TEA e limita o acesso a apoios adequados (Wood-Downie *et al.*, 2021).

Encontramos que a presença de malformação congênita representa cerca de 1.6 a 2.4 vezes mais chances de receber precocemente o diagnóstico de TEA na infância, dado estatístico que permaneceu significativo na análise multivariada. Esse dado está de acordo com outros estudos coortes de nascimento, que indicam que a presença de malformações é um marcador de vulnerabilidade do neurodesenvolvimento (Cleper *et al.*, 2025; Hultman; Sparén; Cnattingius, 2002; Timonen-Soivio *et al.*, 2015). Alinhado a estes achados, há evidências de que malformações urogenitais em homens estão associadas à deficiências nas sinalização de andrógenos, o que pode estar ligada à fisiopatologia do TEA (Wang, Z. *et al.*, 2024), enquanto que em mulheres, as malformações do sistema circulatório parece mais frequente entre aquelas que recebem o diagnóstico de TEA (Cleper *et al.*, 2025; Schieve; Shapira, 2018).

Os fatores maternos (idade dos pais e características sociodemográficas) não obtiveram relevância estatística em nosso estudo, tanto na análise bivariada quanto multivariada mesmo após ajuste LASSO. A idade dos genitores é um fator de risco controverso para o TEA, no entanto, evidências recentes indicam que a idade materna e paterna avançada aumentam independentemente o risco de TEA, mais especificamente para mães acima de 40 anos e pais acima de 50 anos, em comparação com genitores entre 25 e 29 anos (Dehesh *et al.*, 2024; Gao *et al.*, 2020; Lyall *et al.*, 2020; Wu *et al.*, 2017). No entanto, nossa base de dados apresentou uma ausência do preenchimento da idade paterna em mais de 50% das Declarações de

Nascidos Vivos, o que limita o poder estatístico para detectar o efeito real dessa variável, mesmo após as técnicas de imputação (MICE) e regularização (LASSO).

Em nossas análises, o estado civil materno não foi associado ao risco de TEA. Assim como nossos achados, não há evidências sobre a associação direta entre estado civil materno e o risco de autismo (Bahri; Sterrett; Lord, 2023a). Alguns estudos sugerem que famílias de crianças com TEA apresentam risco aumentado de divórcio ao longo do tempo, especialmente em contextos de menor escolaridade materna, o que ocorre como consequência do diagnóstico e dos desafios familiares, não como fator causal (Bahri; Sterrett; Lord, 2023b; Freedman *et al.*, 2012).

Não identificamos associação significativa entre a escolaridade materna e o nosso desfecho, mesmo após ajuste multivariado. Contudo, há evidências de que a baixa escolaridade materna está associada a um maior risco de autismo além de subdiagnóstico e atraso na identificação. (Holland *et al.*, 2026; Liu, X. *et al.*, 2025). Isso pode estar relacionado ao fato de que os nossos dados provêm do setor público, o que cria um perfil sociodemográfico homogêneo na base de dados e mascara o gradiente socioeconômico que pode ser identificado em registros nacionais unificados ou censos autodeclarados.

Quanto a raça/cor, crianças de mães negras, hispânicas ou de grupos minoritários apresentam maior risco de diagnóstico de autismo (Gallin *et al.*, 2025), o que também não foi captado em nosso estudo após os ajustes multivariados. Nossos dados também divergem um pouco dos achados do Censo 2022. A presente divergência pode explicada, novamente, pelo fato de que os nossos dados provêm apenas do setor público, enquanto os dados do Censo refletem o diagnóstico declarado pela população geral, que é fortemente influenciado por desigualdade estruturais e barreiras de acesso aos serviços de saúde (Instituto Brasileiro de Geografia e Estatística, 2025; Tomasiello *et al.*, 2023).

Em relação a paridade, não encontramos associação significativa com o diagnóstico de TEA, mesmo após ajustes multivariados. Não há estudos que detalham diretamente a relação entre número de filhos nascidos vivos ou mortos, indicando que mais pesquisas específicas sobre a paridade e natimortalidade são necessárias. Alguns estudos sugerem que famílias com menos filhos (baixa paridade) tendem a

apresentar maior risco de autismo nos filhos, enquanto famílias com maior número (alta paridade) de filhos apresentam risco reduzido (Cheslack-Postava *et al.*, 2014).

Quanto os fatores gestacionais (número de consultas, duração da gestação, tipo de gravidez e tipo de parto) não encontramos associação significativa com desfecho. Assim como nossos achados, não há evidências atuais que relacione o acompanhamento pré-natal isoladamente ao diagnóstico de autismo (Holland *et al.*, 2026). No entanto, as evidências quanto a duração da gestação permanecem controversas. Outras coortes de nascimento identificaram que a prematuridade e a menor idade gestacional estão associadas a um risco aumentado de TEA, sendo mais elevado para prematuros extremos (Brumbaugh *et al.*, 2020; Ellouk *et al.*, 2025; Wang, L.-W. *et al.*, 2023). Contudo, após ajustes multivariados essa associação é atenuada. Na nossa amostra, houve uma pequena quantidade de prematuros extremos, o que pode ter diluído o efeito estatístico nos modelos multivariados. Em nosso estudo, não identificamos a gemelaridade como um fator de risco para o diagnóstico de autismo. Desta mesma forma, estudos epidemiológicos demonstraram também que a taxa de autismo em gêmeos não difere da taxa de gestações únicas (Hallmayer *et al.*, 2002; Sandin *et al.*, 2014).

Não encontramos, nas análises bivariada e multivariada, relação de tipos de partos com o diagnóstico de TEA. Contudo, há evidências que relacionam a cesariana como fator de risco (OR 1.15 e 1.31) para o TEA, tanto eletivas quanto de emergência (Curran *et al.*, 2015; Liu, K.-Y. *et al.*, 2022). Essa contradição pode ser explicada pelo fato de que o Brasil apresenta uma das maiores taxas de parto operatório do mundo (Pires; Leal; Silva, 2025). Na região de coleta da amostra, há dados que indicam uma taxa superior a 50% de partos operatórios (Tribunal de Contas do Estado de Goiás, 2026), em quanto a OMS preconiza que esse valor seja de aproximadamente 15%. Em um contexto de altíssima prevalência de cesárias, o efeito biológico dessa via de parto pode sofrer um efeito de diluição estatística.

Em nossas análises bivariada e multivariada, o APGAR de 1º e 5º minuto não representaram fatores de risco para o diagnóstico de TEA. No entanto, alguns estudos prévios identificaram que escores baixos de APGAR estão relacionados a um risco aumentado de autismo, especialmente quando o valor do APGAR de 5º minuto é inferior a 7 (He *et al.*, 2022; Modabbernia *et al.*, 2019). Contudo, essa variável é

inconsistente nos estudos e os autores consideram que o APGAR abaixo da faixa de 7-9 deve ser considerado apenas como um fator de risco alarmante, mas não um preditor específico ou suficiente para o diagnóstico de TEA (He *et al.*, 2022; Modabbernia *et al.*, 2019).

De forma análoga, o baixo peso não foi identificado como fator de risco significativo após os ajustes multivariados. De forma contrária a esse achado, múltiplos estudos prévios revelaram que o baixo peso, especialmente o muito baixo peso (< 1500g) está associado a um aumento significativo do risco de TEA, com OR ajustados entre 1.5 e 3.0, mesmo após o controle para outras variáveis perinatais (Dawson; Rieder; Johnson, 2023; Guo *et al.*, 2024; Song *et al.*, 2022). É possível que esse achado em nossa amostra seja um subdiagnóstico artificial de TEA no RAAS-PSI relacionados ao baixo peso, uma vez que os profissionais da saúde tentem a ser mais cautelosos em afirmar um diagnóstico de certeza nesse grupo, preferindo rótulos como “atraso global do desenvolvimento” ou atribuindo os sinais à própria fragilidade neonatal. Além disso, na região da amostra, RNs de baixo peso acabam sendo referenciados para hospitais terciários e/ou programas de seguimento de alto risco, o que impede o registro da produção via RAAS-PSI. Devido a essa possibilidade, também se impõe o viés de Berkson: se a probabilidade de uma criança ir ao CAPSi for menor para aquelas que já estão em tratamento em centros terciários e/ou especializados, a base de dados do CAPSi terá, proporcionalmente menos crianças de baixo peso do que a população geral.

O presente estudo apresenta vários pontos fortes. Primeiro, o pioneirismo no uso da metodologia de *linkage* determinístico para identificar fatores de risco para o TEA em uma coorte de nascidos vivos, utilizando dados de mundo real (RWD) dos sistemas de informação em saúde do Brasil (SINASC e SIA/SUS). Segundo, a robustez estatística é garantida pela aplicação de técnicas avançadas, como a imputação múltipla por equações encadeadas (MICE) para tratar dados ausentes e a regressão penalizada LASSO para a seleção criteriosa de variáveis, o que permitiu mitigar a multicolinearidade e fortalecer a validade dos achados epidemiológicos na região de Goiás. Por fim, o tamanho da amostra, que incluiu o total de 96.112 nascimentos ocorridos em Goiás no ano de 2019, permitindo uma análise populacional ampla e significativa.

Entre as limitações, destaca-se a natureza secundária dos dados, que introduz um potencial viés de classificação, já que o desfecho depende de registros administrativos de diagnóstico em vez de avaliações clínicas diretas. Além disso, a variável idade paterna apresentou uma ausência de preenchimento superior a 50%, o que, apesar das técnicas de imputação, limita o poder estatístico para detectar efeitos reais dessa variável. O estudo também enfrenta o viés de Berkson, uma vez que a amostra restrita aos CAPSi do setor público pode excluir crianças acompanhadas exclusivamente em serviços privados ou centros de alta complexidade, diluindo possíveis associações socioeconômicas e perinatais. Ao realizar um linkage determinístico com dados anônimos, temos como limitação também a falta de identificação exata do par, uma vez que não possuímos variáveis de identificação única entre os bancos de dados utilizados (nome da mãe, por exemplo); o par exato e a certeza no relacionamento entre os bancos de dados só poderiam ser alcançados utilizando dados identificados.

Com base nos achados, recomenda-se a integração de bases de dados da saúde suplementar em pesquisas futuras para capturar um perfil socioeconômico mais diverso e reduzir o efeito de diluição estatística observado no setor público. É fundamental investir no treinamento de profissionais de saúde para melhorar a precisão do preenchimento das declarações de nascidos vivos, especialmente quanto à idade dos genitores e malformações congênitas, visando o fortalecimento das políticas de vigilância. Por fim, sugere-se que os gestores de saúde pública utilizem os indicadores de risco identificados — como o sexo masculino e a presença de malformações — para priorizar o rastreamento e garantir a intervenção precoce nessas populações de maior vulnerabilidade. Além disso, recomenda-se a implementação de estratégias de diagnóstico sensíveis ao gênero, visando reduzir o potencial subdiagnóstico em meninas que utilizam estratégias de "mascaramento", conforme discutido à luz da literatura comparada.

6. CONCLUSÃO

O presente estudo mostrou que incidência acumulada de TEA em crianças de até 5 anos em Goiás é de 1,07%. O sexo masculino e a presença de malformações congênitas foram os principais fatores de risco identificados. O estudo demonstrou que o *linkage* de bases de dados públicas é uma ferramenta viável para gerar evidências nacionais. A homogeneidade socioeconômica do setor público pode ter mascarado outras associações clássicas da literatura. Por fim, os resultados reforçam a necessidade de vigilância precoce nos grupos de maior risco para garantir intervenções oportunas.

REFERÊNCIAS

- ABDELKADER, A.; ALRADINI, F.; ALOSAIMI, A.; ABBAS, A.; JUDEH, Z.; EMY ABU ESAID, T.; SALEH, A.; SHAH, J.; AMER, S. Unveiling the influences of prenatal and maternal factors on the journey of an autistic child. **Frontiers in Psychiatry**, [N.p.], vol. 15, p. 1467821, 20 Dec. 2024. <https://doi.org/10.3389/fpsy.2024.1467821>.
- ALAGHBAND-RAD, J.; HAJIKARIM-HAMEDANI, A.; MOTAMED, M. Camouflage and masking behavior in adult autism. **Frontiers in Psychiatry**, [N.p.], vol. 14, p. 1108110, 2023. <https://doi.org/10.3389/fpsy.2023.1108110>.
- ALMEIDA, M. L.; NEVES, A. S. A Popularização Diagnóstica do Autismo: uma Falsa Epidemia? **Psicologia: Ciência e Profissão**, [N.p.], vol. 40, p. e180896, 2020. <https://doi.org/10.1590/1982-3703003180896>.
- AMERICAN PSYCHIATRIC ASSOCIATION. **Diagnostic and Statistical Manual of Mental Disorders**. DSM-5-TR. [N.p.]: American Psychiatric Association Publishing, 18 Mar. 2022. DOI: 10.1176/appi.books.9780890425787. Available at: <https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425787>. Accessed: 20 Sept. 2025.
- ANDRADE, C. Autism Spectrum Disorder, 1: Genetic and Environmental Risk Factors. **The Journal of Clinical Psychiatry**, [N.p.], vol. 86, no. 2, 14 Apr. 2025. DOI: 10.4088/JCP.25f15878. Available at: <https://www.psychiatrist.com/jcp/autism-spectrum-disorder-1-genetic-and-environmental-risk-factors/>. Accessed: 12 Sept. 2025.
- BAHRI, N.; STERRETT, K.; LORD, C. Marital status over 28 years of parents of individuals with autism and other developmental disabilities. **Journal of Family Psychology**, [N.p.], vol. 37, no. 6, pp. 920–931, Sept. 2023a. <https://doi.org/10.1037/fam0001093>.
- BAHRI, N.; STERRETT, K.; LORD, C. Marital status over 28 years of parents of individuals with autism and other developmental disabilities. **Journal of Family Psychology**, [N.p.], vol. 37, no. 6, pp. 920–931, Sept. 2023b. <https://doi.org/10.1037/fam0001093>.
- BAXTER, A. J.; BRUGHA, T. S.; ERSKINE, H. E.; SCHEURER, R. W.; VOS, T.; SCOTT, J. G. The epidemiology and global burden of autism spectrum disorders. **Psychological Medicine**, [N.p.], vol. 45, no. 3, pp. 601–613, Feb. 2015. <https://doi.org/10.1017/S003329171400172X>.
- BRUMBAUGH, J. E.; WEAVER, A. L.; MYERS, S. M.; VOIGT, R. G.; KATUSIC, S. K. Gestational Age, Perinatal Characteristics, and Autism Spectrum Disorder: A Birth Cohort Study. **The Journal of Pediatrics**, [N.p.], vol. 220, pp. 175–183.e8, May 2020. <https://doi.org/10.1016/j.jpeds.2020.01.022>.
- CHESLACK-POSTAVA, K.; JOKIRANTA, E.; SUOMINEN, A.; LEHTI, V.; SOURANDER, A.; BROWN, A. S. Variation by Diagnostic Subtype in Risk for Autism Spectrum Disorders Associated with Maternal Parity among Finnish Births. **Paediatric and Perinatal Epidemiology**, [N.p.], vol. 28, no. 1, pp. 58–66, Jan. 2014. <https://doi.org/10.1111/ppe.12094>.

CHRISTENSEN, D. L.; MAENNER, M. J.; BILDER, D.; CONSTANTINO, J. N.; DANIELS, J.; DURKIN, M. S.; FITZGERALD, R. T.; KURZIUS-SPENCER, M.; PETTYGROVE, S. D.; ROBINSON, C.; SHENOUDA, J.; WHITE, T.; ZAHORODNY, W.; PAZOL, K.; DIETZ, P. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 4 Years — Early Autism and Developmental Disabilities Monitoring Network, Seven Sites, United States, 2010, 2012, and 2014. **MMWR. Surveillance Summaries**, [N.p.], vol. 68, no. 2, pp. 1–19, 12 Apr. 2019. <https://doi.org/10.15585/mmwr.ss6802a1>.

CLEPER, R.; KAPRA, O.; GOLDENTAL, N.; GROSS, R. The association between autism spectrum disorder and congenital malformations: a population-based nested case-control study. **Molecular Psychiatry**, [N.p.], 15 Oct. 2025. DOI: 10.1038/s41380-025-03302-8. Available at: <https://www.nature.com/articles/s41380-025-03302-8>. Accessed: 8 Feb. 2026.

CURRAN, E. A.; DALMAN, C.; KEARNEY, P. M.; KENNY, L. C.; CRYAN, J. F.; DINAN, T. G.; KHASHAN, A. S. Association Between Obstetric Mode of Delivery and Autism Spectrum Disorder: A Population-Based Sibling Design Study. **JAMA Psychiatry**, [N.p.], vol. 72, no. 9, p. 935, 1 Sept. 2015. <https://doi.org/10.1001/jamapsychiatry.2015.0846>.

DAWSON, G.; RIEDER, A. D.; JOHNSON, M. H. Prediction of autism in infants: progress and challenges. **The Lancet. Neurology**, [N.p.], vol. 22, no. 3, pp. 244–254, Mar. 2023. [https://doi.org/10.1016/S1474-4422\(22\)00407-0](https://doi.org/10.1016/S1474-4422(22)00407-0).

DEHESH, T.; MOSLEH-SHIRAZI, M. A.; JAFARI, S.; ABOLHADI, E.; DEHESH, P. A assessment of the effects of parental age on the development of autism in children: a systematic review and a meta-analysis. **BMC Psychology**, [N.p.], vol. 12, no. 1, p. 685, 22 Nov. 2024. <https://doi.org/10.1186/s40359-024-02184-9>.

DELLAZARI, L.; DE BEM, É. B.; FALCÃO, A. B.; MANJABOSCO, F. D. M.; SORATO, G. B.; BERTO, L. F.; DANTAS, V. A.; DA ROSA, A. L. S. T.; GRAEFF-MARTINS, A. S.; KIELING, R. R.; SALUM JUNIOR, G. A.; ROHDE, L. A.; CAYE, A. Mind the gap: the inconceivable void on the epidemiology of autism spectrum disorders in Brazil. **Brazilian Journal of Psychiatry**, [N.p.], 2024a. DOI: 10.47626/1516-4446-2024-3956. Available at: <https://www.bjp.org.br/details/3585/en-US/mind-the-gap--the-inconceivable-void-on-the-epidemiology-of-autism-spectrum-disorders-in-brazil>. Accessed: 13 Sept. 2025.

DELLAZARI, L.; DE BEM, É. B.; FALCÃO, A. B.; MANJABOSCO, F. D. M.; SORATO, G. B.; BERTO, L. F.; DANTAS, V. A.; DA ROSA, A. L. S. T.; GRAEFF-MARTINS, A. S.; KIELING, R. R.; SALUM JUNIOR, G. A.; ROHDE, L. A.; CAYE, A. Mind the gap: the inconceivable void on the epidemiology of autism spectrum disorders in Brazil. **Brazilian Journal of Psychiatry**, [N.p.], 2024b. DOI: 10.47626/1516-4446-2024-3956. Available at: <https://www.bjp.org.br/details/3585/en-US/mind-the-gap--the-inconceivable-void-on-the-epidemiology-of-autism-spectrum-disorders-in-brazil>. Accessed: 14 Feb. 2026.

DILLON, E. F.; KANNE, S.; LANDA, R. J.; ANNETT, R.; BERNIER, R.; BRADLEY, C.; CARPENTER, L.; KIM, S. H.; PARISH-MORRIS, J.; SCHULTZ, R.; WODKA, E. L.; THE SPARK CONSORTIUM; WODKA, E. L. Sex Differences in Autism: Examining Intrinsic and Extrinsic Factors in Children and Adolescents Enrolled in a

National ASD Cohort. **Journal of Autism and Developmental Disorders**, [N.p.], vol. 53, no. 4, pp. 1305–1318, Apr. 2023. <https://doi.org/10.1007/s10803-021-05385-y>.

DOURADO SOUZA AKAHOSI FERNANDES, A.; VENTURA COUTO, M. C.; COSTA ANDRADA, B.; GODINHO DELGADO, P. G. expansão do diagnóstico de autismo no contexto brasileiro atual: incidência nas políticas públicas e na organização do cuidado. **DESIDADES - Revista Científica da Infância, Adolescência e Juventude**, [N.p.], no. 41, 13 June 2025. DOI: 10.54948/desidades.v1i41.68499. Available at: <https://revistas.ufrj.br/index.php/desidades/article/view/68499>. Accessed: 14 Feb. 2026.

ELLOUK, S.; ZAMSTEIN, O.; WAINSTOCK, T.; SHEINER, E. The association between preterm delivery and autism spectrum disorder in childhood: A retrospective cohort study. **International Journal of Gynecology & Obstetrics**, [N.p.], vol. 170, no. 1, pp. 401–409, July 2025. <https://doi.org/10.1002/ijgo.16150>.

FREEDMAN, B. H.; KALB, L. G.; ZABLOTSKY, B.; STUART, E. A. Relationship Status Among Parents of Children with Autism Spectrum Disorders: A Population-Based Study. **Journal of Autism and Developmental Disorders**, [N.p.], vol. 42, no. 4, pp. 539–548, Apr. 2012. <https://doi.org/10.1007/s10803-011-1269-y>.

GALLIN, Z.; KOLEVZON, A. M.; REICHENBERG, A.; HANKERSON, S. H.; KOLEVZON, A. Racial Differences in the Prevalence of Autism Spectrum Disorder: A Systematic Review. **Journal of Autism and Developmental Disorders**, [N.p.], vol. 55, no. 9, pp. 3364–3377, Sept. 2025. <https://doi.org/10.1007/s10803-024-06403-5>.

GAO, Y.; YU, Y.; XIAO, J.; LUO, J.; ZHANG, Y.; TIAN, Y.; ZHANG, J.; OLSEN, J.; LI, J.; LIEW, Z. Association of Grandparental and Parental Age at Childbirth With Autism Spectrum Disorder in Children. **JAMA Network Open**, [N.p.], vol. 3, no. 4, p. e202868, 15 Apr. 2020. <https://doi.org/10.1001/jamanetworkopen.2020.2868>.

GARDENER, H.; SPIEGELMAN, D.; BUKA, S. L. Prenatal risk factors for autism: comprehensive meta-analysis. **British Journal of Psychiatry**, [N.p.], vol. 195, no. 1, pp. 7–14, July 2009. <https://doi.org/10.1192/bjp.bp.108.051672>.

GUO, B.-Q.; LI, H.-B.; ZHAI, D.-S.; YANG, L.-Q. Prevalence of autism spectrum disorder diagnosis by birth weight, gestational age, and size for gestational age: a systematic review, meta-analysis, and meta-regression. **European Child & Adolescent Psychiatry**, [N.p.], vol. 33, no. 7, pp. 2035–2049, July 2024. <https://doi.org/10.1007/s00787-022-02078-4>.

HALLADAY, A. K.; BISHOP, S.; CONSTANTINO, J. N.; DANIELS, A. M.; KOENIG, K.; PALMER, K.; MESSINGER, D.; PELPHREY, K.; SANDERS, S. J.; SINGER, A. T.; TAYLOR, J. L.; SZATMARI, P. Sex and gender differences in autism spectrum disorder: summarizing evidence gaps and identifying emerging areas of priority. **Molecular Autism**, [N.p.], vol. 6, no. 1, p. 36, Dec. 2015. <https://doi.org/10.1186/s13229-015-0019-y>.

HALLMAYER, J.; GLASSON, E. J.; BOWER, C.; PETTERSON, B.; CROEN, L.; GREYER, J.; RISCH, N. On the Twin Risk in Autism. **The American Journal of Human Genetics**, [N.p.], vol. 71, no. 4, pp. 941–946, Oct. 2002. <https://doi.org/10.1086/342990>.

HE, H.; YU, Y.; WANG, H.; OBEL, C. L.; LI, F.; LI, J. Five-Minute Apgar Score and the Risk of Mental Disorders During the First Four Decades of Life: A Nationwide Registry-Based Cohort Study in Denmark. **Frontiers in Medicine**, [N.p.], vol. 8, p. 796544, 14 Jan. 2022. <https://doi.org/10.3389/fmed.2021.796544>.

HODGES, H.; FEALKO, C.; SOARES, N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. **Translational Pediatrics**, [N.p.], vol. 9, no. S1, pp. S55–S65, Feb. 2020. <https://doi.org/10.21037/tp.2019.09.09>.

HOLLAND, L.; DRUMMOND, K.; THOMSON, S.; SOMINSKY, L.; MARX, W.; LOVE, C.; DAWSON, S. L.; HARRISON, L. C.; SAFFERY, R.; SYMEONIDES, C.; TANG, M. L.; BURGNER, D.; SLY, P. D.; VUILLERMIN, P.; PONSONBY, A.-L.; THE BIS INVESTIGATOR GROUP; MANSELL, T.; RANGANATHAN, S.; O'HELY, M. Prenatal and birth factors associated with child autism diagnosis: a birth cohort perspective. **Pediatric Research**, [N.p.], 15 Jan. 2026. DOI: 10.1038/s41390-025-04747-x. Available at: <https://www.nature.com/articles/s41390-025-04747-x>. Accessed: 14 Feb. 2026.

HULTMAN, C. M.; SPARÉN, P.; CNATTINGIUS, S. Perinatal Risk Factors for Infantile Autism: **Epidemiology**, [N.p.], vol. 13, no. 4, pp. 417–423, July 2002. <https://doi.org/10.1097/00001648-200207000-00009>.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico 2022**. Rio de Janeiro: IBGE, 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico 2022 - Pessoas com deficiência e pessoas diagnosticadas com transtorno do espectro autista**. Rio de Janeiro: IBGE, 2025. 80 pp.

KIM, J. Y.; SON, M. J.; SON, C. Y.; RADUA, J.; EISENHUT, M.; GRESSIER, F.; KOYANAGI, A.; CARVALHO, A. F.; STUBBS, B.; SOLMI, M.; RAIS, T. B.; LEE, K. H.; KRONBICHLER, A.; DRAGIOTI, E.; SHIN, J. I.; FUSAR-POLI, P. Environmental risk factors and biomarkers for autism spectrum disorder: an umbrella review of the evidence. **The Lancet Psychiatry**, [N.p.], vol. 6, no. 7, pp. 590–600, July 2019. [https://doi.org/10.1016/S2215-0366\(19\)30181-6](https://doi.org/10.1016/S2215-0366(19)30181-6).

LAI, M.-C.; BARON-COHEN, S.; BUXBAUM, J. D. Understanding autism in the light of sex/gender. **Molecular Autism**, [N.p.], vol. 6, no. 1, p. 24, Dec. 2015. <https://doi.org/10.1186/s13229-015-0021-4>.

LIU, K.-Y.; TEITLER, J. O.; RAJANANDA, S.; CHEGWIN, V.; BEARMAN, P. S.; HEGYI, T.; REICHMAN, N. E. Elective Deliveries and the Risk of Autism. **American Journal of Preventive Medicine**, [N.p.], vol. 63, no. 1, pp. 68–76, July 2022. <https://doi.org/10.1016/j.amepre.2022.01.024>.

LIU, X.; GUO, C.; LYU, L.; LIU, H.; YU, J.; HUANG, Z.; YANG, L.; WEI, W.; HUANG, J.; LIAO, C.; LI, Y.; CHEN, L.; CHEN, W.-X. Parental educational levels association with autism risk in offsprings and their autistic phenotypes. **Psychiatry Research**, [N.p.], vol. 351, p. 116649, Sept. 2025. <https://doi.org/10.1016/j.psychres.2025.116649>.

LOCKWOOD ESTRIN, G.; MILNER, V.; SPAIN, D.; HAPPÉ, F.; COLVERT, E. Barriers to Autism Spectrum Disorder Diagnosis for Young Women and Girls: a

Systematic Review. **Review Journal of Autism and Developmental Disorders**, [N.p.], vol. 8, no. 4, pp. 454–470, 2021. <https://doi.org/10.1007/s40489-020-00225-8>.

LOOMES, R.; HULL, L.; MANDY, W. P. L. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. **Journal of the American Academy of Child and Adolescent Psychiatry**, [N.p.], vol. 56, no. 6, pp. 466–474, June 2017a. <https://doi.org/10.1016/j.jaac.2017.03.013>.

LOOMES, R.; HULL, L.; MANDY, W. P. L. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. **Journal of the American Academy of Child & Adolescent Psychiatry**, [N.p.], vol. 56, no. 6, pp. 466–474, June 2017b. <https://doi.org/10.1016/j.jaac.2017.03.013>.

LYALL, K.; SONG, L.; BOTTERON, K.; CROEN, L. A.; DAGER, S. R.; FALLIN, M. D.; HAZLETT, H. C.; KAUFFMAN, E.; LANDA, R.; LADD-ACOSTA, C.; MESSINGER, D. S.; OZONOFF, S.; PANDEY, J.; PIVEN, J.; SCHMIDT, R. J.; SCHULTZ, R. T.; STONE, W. L.; NEWSCHAFER, C. J.; VOLK, H. E. The Association Between Parental Age and Autism-Related Outcomes in Children at High Familial Risk for Autism. **Autism Research: Official Journal of the International Society for Autism Research**, [N.p.], vol. 13, no. 6, pp. 998–1010, June 2020. <https://doi.org/10.1002/aur.2303>.

MAENNER, M. J.; WARREN, Z.; WILLIAMS, A. R.; AMOAKOHE, E.; BAKIAN, A. V.; BILDER, D. A.; DURKIN, M. S.; FITZGERALD, R. T.; FURNIER, S. M.; HUGHES, M. M.; LADD-ACOSTA, C. M.; MCARTHUR, D.; PAS, E. T.; SALINAS, A.; VEHORN, A.; WILLIAMS, S.; ESLER, A.; GRZYBOWSKI, A.; HALL-LANDE, J.; NGUYEN, R. H. N.; PIERCE, K.; ZAHORODNY, W.; HUDSON, A.; HALLAS, L.; MANCILLA, K. C.; PATRICK, M.; SHENOUDA, J.; SIDWELL, K.; DIRIENZO, M.; GUTIERREZ, J.; SPIVEY, M. H.; LOPEZ, M.; PETTYGROVE, S.; SCHWENK, Y. D.; WASHINGTON, A.; SHAW, K. A. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. **Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002)**, [N.p.], vol. 72, no. 2, pp. 1–14, 24 Mar. 2023. <https://doi.org/10.15585/mmwr.ss7202a1>.

MANDY, W.; LAI, M.-C. Towards sex- and gender-informed autism research. **Autism**, [N.p.], vol. 21, no. 6, pp. 643–645, Aug. 2017. <https://doi.org/10.1177/1362361317706904>.

MODABBERNIA, A.; SANDIN, S.; GROSS, R.; LEONARD, H.; GISSLER, M.; PARNER, E. T.; FRANCIS, R.; CARTER, K.; BRESNAHAN, M.; SCHENDEL, D.; HORNIG, M.; REICHENBERG, A. Apgar score and risk of autism. **European Journal of Epidemiology**, [N.p.], vol. 34, no. 2, pp. 105–114, Feb. 2019. <https://doi.org/10.1007/s10654-018-0445-1>.

PAULA, C. S.; RIBEIRO, S. H.; FOMBONNE, E.; MERCADANTE, M. T. Brief report: prevalence of pervasive developmental disorder in Brazil: a pilot study. **Journal of Autism and Developmental Disorders**, [N.p.], vol. 41, no. 12, pp. 1738–1742, Dec. 2011. <https://doi.org/10.1007/s10803-011-1200-6>.

PIRES, R. C. R.; LEAL, M. D. C.; SILVA, A. A. M. D. Recent temporal trends in cesarean section rates in Brazil according to the Robson classification system. **Cadernos de Saúde Pública**, [N.p.], vol. 41, no. 8, p. e00154024, 2025. <https://doi.org/10.1590/0102-311xen154024>.

QIU, S.; QIU, Y.; LI, Y.; CONG, X. Genetics of autism spectrum disorder: an umbrella review of systematic reviews and meta-analyses. **Translational Psychiatry**, [N.p.], vol. 12, no. 1, p. 249, 15 June 2022. <https://doi.org/10.1038/s41398-022-02009-6>.

RUSSELL, G.; STAPLEY, S.; NEWLOVE-DELGADO, T.; SALMON, A.; WHITE, R.; WARREN, F.; PEARSON, A.; FORD, T. Time trends in autism diagnosis over 20 years: a UK population-based cohort study. **Journal of Child Psychology and Psychiatry, and Allied Disciplines**, [N.p.], vol. 63, no. 6, pp. 674–682, June 2022. <https://doi.org/10.1111/jcpp.13505>.

SANDIN, S.; LICHTENSTEIN, P.; KUJA-HALKOLA, R.; LARSSON, H.; HULTMAN, C. M.; REICHENBERG, A. The Familial Risk of Autism. **JAMA**, [N.p.], vol. 311, no. 17, p. 1770, 7 May 2014. <https://doi.org/10.1001/jama.2014.4144>.

SCHIEVE, L. A.; SHAPIRA, S. K. Invited Commentary: Male Reproductive System Congenital Malformations and the Risk of Autism Spectrum Disorder. **American Journal of Epidemiology**, [N.p.], vol. 187, no. 4, pp. 664–667, 1 Apr. 2018. <https://doi.org/10.1093/aje/kwx369>.

SHAW, K. A.; WILLIAMS, S.; PATRICK, M. E.; VALENCIA-PRADO, M.; DURKIN, M. S.; HOWERTON, E. M.; LADD-ACOSTA, C. M.; PAS, E. T.; BAKIAN, A. V.; BARTHOLOMEW, P.; NIEVES-MUÑOZ, N.; SIDWELL, K.; ALFORD, A.; BILDER, D. A.; DIRIENZO, M.; FITZGERALD, R. T.; FURNIER, S. M.; HUDSON, A. E.; POKOSKI, O. M.; SHEA, L.; TINKER, S. C.; WARREN, Z.; ZAHORODNY, W.; AGOSTO-ROSA, H.; ANBAR, J.; CHAVEZ, K. Y.; ESLER, A.; FORKNER, A.; GRZYBOWSKI, A.; AGIB, A. H.; HALLAS, L.; LOPEZ, M.; MAGAÑA, S.; NGUYEN, R. H. N.; PARKER, J.; PIERCE, K.; PROTHO, T.; TORRES, H.; VANEGAS, S. B.; VEHORN, A.; ZHANG, M.; ANDREWS, J.; GREER, F.; HALL-LANDE, J.; MCARTHUR, D.; MITAMURA, M.; MONTES, A. J.; PETTYGROVE, S.; SHENOUDA, J.; SKOWYRA, C.; WASHINGTON, A.; MAENNER, M. J. Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years — Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022. **MMWR. Surveillance Summaries**, [N.p.], vol. 74, no. 2, pp. 1–22, 17 Apr. 2025. <https://doi.org/10.15585/mmwr.ss7402a1>.

SONG, I. G.; KIM, H.-S.; CHO, Y.-M.; LIM, Y.; MOON, D.-S.; SHIN, S. H.; KIM, E.-K.; PARK, J.; SHIN, J. E.; HAN, J.; EUN, H. S. Association between birth weight and neurodevelopmental disorders assessed using the Korean National Health Insurance Service claims data. **Scientific Reports**, [N.p.], vol. 12, no. 1, p. 2080, 8 Feb. 2022. <https://doi.org/10.1038/s41598-022-06094-x>.

SOUSAMLI, A.; DRAGIOTI, E.; METALLINO, D.; LYKERIDOU, A.; DOUROU, P.; ATHANASIADOU, C. R.; ANAGNOSTOPOULOS, D.; SARANTAKI, A. Perinatal and Demographic Risk Factors Associated with Autism Spectrum Disorder: A National Survey of Potential Predictors and Severity. **Healthcare**, [N.p.], vol. 12, no. 20, p. 2057, 16 Oct. 2024. <https://doi.org/10.3390/healthcare12202057>.

SUKIENNIK, R.; MARCHEZAN, J.; SCORNAVACCA, F. Challenges on Diagnoses and Assessments Related to Autism Spectrum Disorder in Brazil: A Systematic Review. **Frontiers in Neurology**, [N.p.], vol. 12, p. 598073, 20 Jan. 2022. <https://doi.org/10.3389/fneur.2021.598073>.

TIMONEN-SOIVIO, L.; VANHALA, R.; MALM, H.; LEIVONEN, S.; JOKIRANTA, E.; HINKKA-YLI-SALOMÄKI, S.; GISSLER, M.; BROWN, A. S.; SOURANDER, A. The

association between congenital anomalies and autism spectrum disorders in a Finnish national birth cohort. **Developmental Medicine & Child Neurology**, [N.p.], vol. 57, no. 1, pp. 75–80, Jan. 2015. <https://doi.org/10.1111/dmcn.12581>.

TOMASIELLO, D. B.; BAZZO, J.; PARGA, J.; SERVO, L. M.; PEREIRA, R. H. M. TD 2832 - Desigualdades raciais e de renda no acesso à saúde nas cidades brasileiras. **Texto para Discussão**, [N.p.], pp. 1–38, 11 Jan. 2023. <https://doi.org/10.38116/td2832>.

TRIBUNAL DE CONTAS DO ESTADO DE GOIÁS. **Partos Cesáreos**. [N.p.]: [N.p.], [N.d.]. Available at: <https://portal.tce.go.gov.br/partos-cesareos>.

WAIZBARD-BARTOV, E.; FEIN, D.; LORD, C.; AMARAL, D. G. Autism severity and its relationship to disability. **Autism Research**, [N.p.], vol. 16, no. 4, pp. 685–696, Apr. 2023. <https://doi.org/10.1002/aur.2898>.

WANG, C.; GENG, H.; LIU, W.; ZHANG, G. Prenatal, perinatal, and postnatal factors associated with autism: A meta-analysis. **Medicine**, [N.p.], vol. 96, no. 18, p. e6696, May 2017. <https://doi.org/10.1097/MD.00000000000006696>.

WANG, L.-W.; LIN, H.-C.; TSAI, M.-L.; CHANG, Y.-T.; CHANG, Y.-C. Preterm birth and small for gestational age potentiate the association between maternal hypertensive pregnancy and childhood autism spectrum disorder. **Scientific Reports**, [N.p.], vol. 13, no. 1, p. 9606, 13 June 2023. <https://doi.org/10.1038/s41598-023-36787-w>.

WANG, Z.; ZHANG, B.; MU, C.; QIAO, D.; CHEN, H.; ZHAO, Y.; CUI, H.; ZHANG, R.; LI, S. Androgen levels in autism spectrum disorders: a systematic review and meta-analysis. **Frontiers in Endocrinology**, [N.p.], vol. 15, p. 1371148, 2024. <https://doi.org/10.3389/fendo.2024.1371148>.

WOOD-DOWNIE, H.; WONG, B.; KOVSHOFF, H.; MANDY, W.; HULL, L.; HADWIN, J. A. Sex/Gender Differences in Camouflaging in Children and Adolescents with Autism. **Journal of Autism and Developmental Disorders**, [N.p.], vol. 51, no. 4, pp. 1353–1364, Apr. 2021. <https://doi.org/10.1007/s10803-020-04615-z>.

WU, S.; WU, F.; DING, Y.; HOU, J.; BI, J.; ZHANG, Z. Advanced parental age and autism risk in children: a systematic review and meta-analysis. **Acta Psychiatrica Scandinavica**, [N.p.], vol. 135, no. 1, pp. 29–41, Jan. 2017. <https://doi.org/10.1111/acps.12666>.

YANG, B.; ZAKS, N.; KAJANTIE, E.; PERSSON, M. S. M.; REICHENBERG, A.; GISSLER, M.; RISNES, K.; KOLEVZON, A.; ÅDÉN, U.; SUSSER, E.; PERSSON, M.; LUDVIGSSON, J. F.; TAMMIMIES, K.; POON, L. C.; YIP, B.; DÖRING, N.; SANDIN, S.; YIN, W. Risk Factors for Autism Spectrum Disorder in Individuals Born Preterm: A Systematic Review and Meta-Analysis of Population-Based Studies. **Biological Psychiatry Global Open Science**, [N.p.], vol. 5, no. 5, p. 100535, Sept. 2025. <https://doi.org/10.1016/j.bpsgos.2025.100535>.

ZEIDAN, J.; FOMBONNE, E.; SCORAH, J.; IBRAHIM, A.; DURKIN, M. S.; SAXENA, S.; YUSUF, A.; SHIH, A.; ELSABBAGH, M. Global prevalence of autism: A systematic review update. **Autism Research: Official Journal of the International Society for Autism Research**, [N.p.], vol. 15, no. 5, pp. 778–790, May 2022. <https://doi.org/10.1002/aur.2696>.

ANEXOS – LINHAS DE CÓDIGO EM R UTILIZADAS

1. Linkage

```
#####
## CONFIGURAÇÃO ####
#####

library(pacman)

pacman::p_load(read.dbc, foreign, dplyr, tidyverse, reclin2,
               digest, knitr, DT, mice, gtsummary)

#####
## BANCO DE DADOS ####
#####

# Define as variáveis de blocagem
VARIAVEL_BLOCAGEM <- "DTNASC"

# Define as variáveis de comparação
VARIABLEIS_COMPARACAO <- c("DTNASC", "SEXO", "RACACOR")

# Define a pontuação mínima para ser considerado um par
LIMITE_SCORE <- 0.5

# Banco de Dados do SINASC
SINASC <- read.csv("E:/TC_DB/SINASC/DNGO_FILTRO.csv",
                  stringsAsFactors = F, fileEncoding = "latin1", sep = ";")

# Banco de Dados do RAAS
RAAS <- read.csv("E:/TC_DB/SIA/PSGO_PROCESSADO.csv",
                 stringsAsFactors = F, fileEncoding = "latin1", sep = ";")

#####
## PADRONIZAÇÃO ####
#####

# Alterando os valores na coluna
SINASC <- SINASC %>%
  mutate(SEXO = case_when(
    SEXO == "1" ~ "M",
    SEXO == "2" ~ "F",
  ))

# Renomeando a coluna
RAAS <- RAAS %>%
  rename(SEXO = SEXOPAC)

#####
## LINKAGE V2 ####
#####

# Cria os pares e calcula os scores
pares_com_scores <- pair_blocking(SINASC, RAAS, VARIAVEL_BLOCAGEM) %>%
  compare_pairs(VARIABLEIS_COMPARACAO, default_comparator = cmp_identical()) %>%
  {
    modelo <- problink_em(as.formula(paste("~", paste(VARIABLEIS_COMPARACAO, collapse =
"+"))), data = .)
    predict(modelo, pairs = ., add = TRUE)
  }

# Passo de Diagnóstico
print(summary(pares_com_scores$weights))
```

```

hist(pares_com_scores$weights, breaks = 50, main = "Distribuição dos Scores do Linkage")

# Com base no diagnóstico, prossiga com o linkage
linkage_final2 <- pares_com_scores %>%
  select_threshold("threshold", score = "weights", threshold = LIMITE_SCORE) %>%
  select_greedy("weights", variable = "greedy", threshold = 0) %>%
  select_n_to_m("weights", variable = "ntom", threshold = 0) %>%
  link(selection = "ntom")

# Visualiza o resultado
head(linkage_final)

#####
## VERIFICAR DUPLICATAS ###
#####

# Contar quantas vezes cada 'CONTADOR' aparece
contagem_duplicatas <- linkage_final2 %>%
  group_by(CONTADOR) %>%
  summarise(n = n()) %>%
  filter(n > 1) %>%
  arrange(desc(n))

# Ver a contagem
print(contagem_duplicatas)
# Isso mostrará quais 'CONTADOR' estão repetidos e quantas vezes

total_registros_duplicados <- sum(contagem_duplicatas$n - 1)
print(paste("Número total de registros duplicados a serem removidos:",
total_registros_duplicados))

#####
## FILTRO < 1 ANOS E CALCULO DA IDADE ###
#####

banco_filtrado_idade <- banco_com_desfecho %>%
  mutate(
    # --- Corrigir DTNASC ---
    # Converte para string (caso seja número)
    DTNASC_str = as.character(DTNASC.x),

    # Adiciona o "0" à esquerda se a data tiver 7 dígitos (ex: 9102019)
    DTNASC_pad = str_pad(DTNASC_str, width = 8, side = "left", pad = "0"),

    # Converte a string corrigida para um objeto Data
    DTNASC_date = as.Date(DTNASC_pad, format = "%d%m%Y"),

    # --- Corrigir DT_ATEND ---
    # Converte para string
    DT_ATEND_str = as.character(DT_ATEND),

    # Adiciona o dia "01" ao final da string YYYYMM
    DT_ATEND_full = paste0(DT_ATEND_str, "01"),

    # Converte a string "YYYYMMDD" para um objeto Data
    DT_ATEND_date = as.Date(DT_ATEND_full, format = "%Y%m%d"),

    # --- Calcular Idade ---
    # Calcula o intervalo em meses entre o nascimento e o atendimento
    # O operador %/% meses(1) dá a idade em meses completos
    idade_meses = interval(DTNASC_date, DT_ATEND_date) %/% months(1)
  ) %>%

# Mantém apenas crianças com 12 meses ou mais (>= 1 ano)
filter(idade_meses >= 12)

```

```

# Verifica o resultado
print(paste("Registros antes do filtro de idade:", nrow(banco_com_desfecho)))
print(paste("Registros após remover menores de 1 ano:", nrow(banco_filtrado_idade)))

# (Opcional) Ver a distribuição de idade dos removidos
# banco_removidos <- anti_join(banco_com_desfecho, banco_filtrado_idade, by = "CONTADOR")
# summary(banco_removidos$idade_meses)

write.csv2(banco_filtrado_idade, file = "CASOS_FILTRADO.csv")

#####
## CASOS ###
#####

banco_final <- readr::read_csv2("BANCOFINAL.csv")

# Define a lista de CIDs de interesse
lista_cid_tea <- c("F840", "F841", "F845", "F848", "F849", "F84")

# Define o Banco de Dados com o desfecho de interesse dado pelo CID
banco_com_desfecho <- banco_final %>%
  mutate(
    # Cria uma coluna CIDPRI limpa: remove pontos, espaços e deixa maiúscula.
    CIDPRI_limpo = str_replace_all(toupper(CIDPRI), "[[:punct:]]|[[:space:]]", ""),

    # Cria a variável 'desfecho'
    desfecho = ifelse(CIDPRI_limpo %in% lista_cid_tea, 1, 0)
  )

# Verifica o resultado
print("Contagem de Casos (1) vs. Não-Casos (0) nos registros únicos:")
table(banco_com_desfecho$desfecho)

```

2. Formatação do banco de dados

```
#####
## CONFIGURAÇÃO ####
#####

library(pacman)

pacman::p_load(read.dbc, foreign, dplyr, tidyverse, reclin2,
                digest, knitr, DT, mice)

#####
## FORMATAÇÃO DAS VARIÁVEIS ####
#####

# Prepara as variáveis para imputação conforme Quadros 1 e 2
banco_para_imputar <- readr::read_csv2("BANCOFINALV2.csv") %>%
  mutate(

    # --- Variáveis de Desfecho ---
    desfecho = as.factor(desfecho), # 0=Controle, 1=Caso

    # --- Quadro 1: Características Maternas ---

    # IDADEMAE: Categórica (< 20, 20-34, 35+)
    IDADEMAE_cat = case_when(
      IDADEMAE < 30 ~ "<30",
      IDADEMAE >= 30 & IDADEMAE <= 34 ~ "30-34",
      IDADEMAE >= 35 ~ "35+",
      TRUE ~ NA_character_
    ),

    # ESCMAE: Categórica (3 ou menos, 4 a 7, 8 a 11 anos)
    ESCMAE_cat = case_when(
      ESCMAE <= 2 ~ "0-3 anos",
      ESCMAE >= 3 & ESCMAE < 4 ~ "4-7 anos",
      ESCMAE >= 4 & ESCMAE <= 5 ~ "8-11 anos",
      TRUE ~ NA_character_
    ),

    # ESTCIVMAE: Categórica (Com companheiro, Sem companheiro)
    ESTCIVMAE_cat = case_when(
      ESTCIVMAE %in% c(2, 5) ~ "Com companheiro",
      ESTCIVMAE %in% c(1, 3, 4, 9) ~ "Sem companheiro",
      TRUE ~ NA_character_
    ),

    # RACACORMAE: Categórica (Branca, Preta/Parda, Amarela, Indígena)
    RACACORMAE_cat = case_when(
      RACACORMAE == 1 ~ "Branca",
      RACACORMAE %in% c(2, 3, 4, 5) ~ "Preta/Parda/Amarela/Indígena",
      TRUE ~ NA_character_
    ),

    # QTDFILVIVO: Categórica (0, 1, 2+)
    QTDFILVIVO_cat = case_when(
      QTDFILVIVO == 0 ~ "0",
      QTDFILVIVO == 1 ~ "1 ou mais",
      QTDFILVIVO >= 2 ~ "1 ou mais",
      TRUE ~ NA_character_
    ),

    # QTDFILMORT: Categórica (0, 1, 2+)
    QTDFILMORT_cat = case_when(
      QTDFILMORT == 0 ~ "0",
```

```

    QTDFILMORT == 1 ~ "1 ou mais",
    QTDFILMORT >= 2 ~ "1 ou mais",
    TRUE ~ NA_character_
  ),

# GESTACAO: Categórica (<37 pré-termo; >=37 termo)
GESTACAO_cat = case_when(
  GESTACAO %in% c(1, 2, 3, 4) ~ "Pré-termo (<37)",
  GESTACAO %in% c(5, 6) ~ "Termo (>=37)",
  TRUE ~ NA_character_
),

# CONSULTAS: Categórica (0-6 Inadequado, >=7 Adequado)
CONSULTAS_cat = case_when(
  CONSULTAS %in% c(1, 2, 3) ~ "Inadequado (0-6)",
  CONSULTAS == 4 ~ "Adequado (>=7)",
  TRUE ~ NA_character_
),

# PARTO: Dicotômica (Vaginal, Operatório)
PARTO_cat = case_when(
  PARTO == 1 ~ "Vaginal",
  PARTO == 2 ~ "Operatório (Cesárea)",
  TRUE ~ NA_character_
),

# IDADEPAI: Categórica (< 20, 20-34, 35+)
# Esta é a variável com 52% de NAs que o MICE irá imputar
IDADEPAI_cat = case_when(
  IDADEPAI < 30 ~ "<30",
  IDADEPAI >= 30 & IDADEPAI <= 34 ~ "30-34",
  IDADEPAI >= 35 & IDADEPAI < 40 ~ "35-40",
  IDADEPAI >= 40 ~ "40+",
  TRUE ~ NA_character_
),

# --- Quadro 2: Características do RN ---

# SEXO: Dicotômica (Masculino, Feminino)
SEXO_cat = case_when(
  SEXO.x == "M" ~ "Masculino",
  SEXO.x == "F" ~ "Feminino",
  TRUE ~ NA_character_
),

# PESO: Categórica (<2500g e >=2500g)
PESO_cat = case_when(
  PESO < 2500 ~ "Baixo peso (<2500g)",
  PESO >= 2500 ~ "Peso normal (>=2500g)",
  TRUE ~ NA_character_
),

# APGAR1: Categórica (<=7 e >7)
APGAR1_cat = case_when(
  APGAR1 <= 7 ~ "<=7",
  APGAR1 > 7 ~ ">7",
  TRUE ~ NA_character_
),

# APGAR5: Categórica (<=7 e >7)
APGAR5_cat = case_when(
  APGAR5 <= 7 ~ "<=7",
  APGAR5 > 7 ~ ">7",
  TRUE ~ NA_character_
),

```

```

# IDANOMAL: Dicotômica (Sim, Não)
IDANOMAL_cat = case_when(
  IDANOMAL == 1 ~ "Sim",
  IDANOMAL == 2 ~ "Não",
  IDANOMAL == 9 ~ NA_character_, # Trata "Ignorado" como NA
  TRUE ~ NA_character_
),

# GRAVIDEZ: Categórica (Única, Múltipla)
GRAVIDEZ_cat = case_when(
  GRAVIDEZ == 1 ~ "Única",
  GRAVIDEZ %in% c(2, 3) ~ "Múltipla", # Agrupa 2 (Dupla) e 3 (Tripla+)
  GRAVIDEZ == 9 ~ NA_character_, # Trata "Ignorado" como NA
  TRUE ~ NA_character_
)
)%>%

# Converter todas as novas variáveis _cat para 'factor'
# O 'across' aplica a mesma função (as.factor) em todas as colunas
# que terminam com "_cat"
mutate(across(ends_with("_cat"), as.factor)) %>%

# --- DEFINIR AS CATEGORIAS DE REFERÊNCIA ---
# O 'ref' deve ser o grupo de "não exposto" ou "menor risco"
mutate(
  # Fator de Risco: Idade avançada
  IDADEMAE_cat = relevel(IDADEMAE_cat, ref = "30-34"),
  IDADEPAI_cat = relevel(IDADEPAI_cat, ref = "30-34"), #

  # Fator de Risco: Baixa escolaridade
  ESCMAE_cat = relevel(ESMAE_cat, ref = "8-11 anos"),

  # Fator de Risco: Prematuridade
  GESTACAO_cat = relevel(GESTACAO_cat, ref = "Termo (>=37)"),

  # Fator de Risco: Cuidado inadequado
  CONSULTAS_cat = relevel(CONSULTAS_cat, ref = "Adequado (>=7)"),

  # Fator de Risco: Parto cesáreo
  PARTO_cat = relevel(PARTO_cat, ref = "Vaginal"),

  # Fator de Risco: Sexo masculino
  SEXO_cat = relevel(SEXO_cat, ref = "Feminino"),

  # Fator de Risco: Baixo peso
  PESO_cat = relevel(PESO_cat, ref = "Peso normal (>=2500g)'),

  # Fator de Risco: Apgar <= 7
  APGAR1_cat = relevel(APGAR1_cat, ref = ">7"),
  APGAR5_cat = relevel(APGAR5_cat, ref = ">7"),

  # Fator de Risco: Anomalia Sim
  IDANOMAL_cat = relevel(IDANOMAL_cat, ref = "Não"),

  # Fator de Risco: Múltipla
  GRAVIDEZ_cat = relevel(GRAVIDEZ_cat, ref = "Única"),

  # Fator de Risco: Sem companheiro
  ESTCIVMAE_cat = relevel(ESTCIVMAE_cat, ref = "Com companheiro"),

  # Desfecho: TEA
  desfecho = relevel(desfecho, ref = "0")
)

```

3. MICE e Análise Univariada/Multivariada

```
#####
## VERIFICAR MCAR ###
#####

# --- Prepara o banco de dados para o MCAR ---
banco_para_tabela <- readr::read_csv2("BANCOFINAL.csv") %>%
  mutate(
    IDADEPAI_ausente = ifelse(is.na(IDADEPAI), "Ausente", "Preenchido"),
    PESO_ausente = ifelse(is.na(PESO), "Ausente", "Preenchido"),
    IDANOMAL_ausente = ifelse(is.na(IDANOMAL), "Ausente", "Preenchido"),
    ESTCIVMAE_ausente = ifelse(is.na(ESTCIVMAE), "Ausente", "Preenchido"),
    ESCMAE_ausente = ifelse(is.na(ESMAE), "Ausente", "Preenchido"),
  )

# Cria uma tabela comparando Casos vs. Controles
# "Tabela 1" de avaliação de dados ausentes
tabela_ausentes <- banco_para_tabela %>%
  select(desfecho, IDADEPAI_ausente, PESO_ausente, IDANOMAL_ausente) %>%
  tbl_summary(
    by = desfecho, # Compara Casos (1) vs Controles (0)
    percent = "column",
    label = list(
      IDADEPAI_ausente ~ "Idade do Pai"
    )
  ) %>%
  add_p() # Adiciona o p-valor

print(tabela_ausentes)

#####
## MICE ###
#####

# --- Prepara o banco de dados para o MICE ---

banco_para_imputar <- banco_com_desfecho %>%
  mutate(
    # Converter o desfecho para fator (crucial para o MICE)
    desfecho = as.factor(desfecho),

    # Converter todas as variáveis categóricas para fator
    ESCMAE = as.factor(ESMAE),
    ESTCIVMAE = as.factor(ESTCIVMAE),
    RACACORMAE = as.factor(RACACORMAE),
    PARTO = as.factor(PARTO),
    GESTACAO = as.factor(GESTACAO),
    IDANOMAL = as.factor(IDANOMAL),
    SEXO = as.factor(SEXO),
    APGAR1 = as.factor(APGAR1),
    APGAR5 = as.factor(APGAR5)
  )

# Selecionar APENAS as variáveis que você usará no seu modelo final
# O MICE usará todas elas para prever umas às outras
variaveis_do_modelo <- c(
  "desfecho", "IDADEMAE", "IDADEPAI", "ESMAE", "PARTO", "PESO",
  "GESTACAO", "APGAR5", "IDANOMAL", "CONSULTAS", "RACACORMAE"
) # Adicione todas as suas variáveis de interesse aqui

banco_subset <- banco_para_imputar[, variaveis_do_modelo]

# --- IMPUTAR OS DADOS ---
```

```

# 1.1. Executar a imputação
# m=5 significa que ele criará 5 bancos de dados completos.
# É um número comum para testes. Para o estudo atual, 5 ou 10 está bom.
# 'maxit=5' é o número de iterações (para rodar mais rápido)
# O MICE irá imputar automaticamente IDADEPAI, RACACORMAE e
# todos os outros pequenos NAs (ESMAE, PARTO, etc.)
imputacao_mice <- mice(banco_subset, m = 5, maxit = 5, method = 'pmm', seed = 123)

# 'pmm' (predictive mean matching) é um bom método padrão para
# variáveis contínuas como IDADEPAI. O MICE escolherá os métodos
# corretos para os seus fatores (como RACACORMAE) automaticamente.

#####
## CONFIGURAÇÃO ####
#####

library(pacman)

pacman::p_load(read.dbc, foreign, dplyr, tidyverse, reclin2,
               digest, knitr, DT, mice)

#####
## MODELO UNIVARIADO ####
#####

# Variáveis categóricas que acabamos de criar
#"ESMAE_cat", "ESTCIVMAE_cat", "RACACORMAE_cat", "QTDFILVIVO_cat",
#"QTDFILMORT_cat", "GESTACAO_cat", "PARTO_cat", "SEXO_cat",
#"APGAR1_cat", "APGAR5_cat", "IDANOMAL_cat", "PESO_cat", "GRAVIDEZ_cat",
#"IDADEMAE_cat", "IDADEPAI_cat", "CONSULTAS_cat"

# --- Ajustar o Modelo Univariado ---
modelos_univariados <- with(
  imputacao_mice,
  glm(desfecho ~ APGAR1_cat, family = "binomial")
)

# --- Combinar ("Poolar") os Resultados ---
# Combina os 5 resultados em um só
resultados_poolados <- pool(modelos_univariados)

# --- Extrair os Odds Ratios ---
# A opção 'exponentiate = TRUE' converte os log-odds em Odds Ratios (OR)
summary_ors <- summary(
  resultados_poolados,
  exponentiate = TRUE,
  conf.int = 0.95
)

# --- Ver a Tabela de Resultados ---
print(summary_ors, digits = 4)

lista_preditores <- c(
  "ESMAE_cat", "ESTCIVMAE_cat", "RACACORMAE_cat", "IDADEMAE_cat", "IDADEPAI_cat",
  "QTDFILVIVO_cat", "QTDFILMORT_cat", "CONSULTAS_cat", "GESTACAO_cat", "GRAVIDEZ_cat",
  "PARTO_cat", "APGAR1_cat", "APGAR5_cat", "SEXO_cat", "PESO_cat", "IDANOMAL_cat"
)

# --- Criar o modelo nulo ---
modelo_nulo <- with(
  imputacao_mice,
  glm(as.formula("desfecho ~ 1"), family = "binomial")
)

# --- Automatizar a análise com lapply ---

```

```

# 'lapply' irá executar o código abaixo para CADA item da 'lista_preditores'

lista_de_resultados <- lapply(lista_preditores, function(var_name) {

  modelos_uni <- with(
    imputacao_mice,
    glm(as.formula(paste("desfecho ~", var_name)), family = "binomial")
  )

  # 3. Combina ("poola") os 5 modelos
  modelos_poolados <- pool(modelos_uni)

  # 4. Obtém o P-VALOR GLOBAL (para Tabela 1)
  # Compara o modelo da variável com o modelo nulo
  teste_global <- D3(modelos_uni, modelo_nulo)
  p_valor_global <- teste_global$p.value

  # 5. Obtém os ORs das CATEGORIAS (para Tabela 2)
  sumario_categorias <- summary(modelos_poolados, exponentiate = TRUE, conf.int = 0.95)

  # 6. Limpa e formata o resultado
  sumario_df <- as.data.frame(sumario_categorias)
  sumario_df <- sumario_df[sumario_df$term != "(Intercept)", ] # Remove o intercepto
  sumario_df$p_valor_global_da_variavel <- p_valor_global

  return(sumario_df)
})

# --- Combinar tudo em uma Tabela Final ---
# Converte a lista de resultados em um único data.frame
tabela_univariada_final <- dplyr::bind_rows(lista_de_resultados)

# Reordena as colunas para melhor visualização
tabela_univariada_final <- tabela_univariada_final %>%
  select(
    term,
    p.value,
    estimate,
    p.value,
    `2.5 %`,
    `97.5 %`
  )

# --- Ver a Tabela Final ---
# Desabilita notação científica e ajusta dígitos
options(scipen = 999)

# Use print() com digits para formatar as casas decimais
print(tabela_univariada_final, digits = 3)

#####
## MODELO MULTIVARIADO ####
#####

# --- BLOCO 1 ---

vars_bloco1 <- "ESTCIVMAE_cat + QTDFILMORT_cat + IDADEMAE_cat + IDADEPAI_cat"

formula_str_1 <- paste("desfecho ~", vars_bloco1)

# --- EXECUTAR A ANÁLISE (GLM nos dados imputados) ---
# Executar o modelo final nos 5 bancos de dados imputados

modelos_bloco_1 <- with(
  imputacao_mice,
  glm(as.formula(formula_str_1), family = "binomial")
)

```

```

)

# --- COMBINAR OS RESULTADOS ("POOL") ---
# Combinar (pool) os resultados

pool_bloco_1 <- pool(modelos_bloco_1)
print(summary(pool_bloco_1, exponentiate = TRUE, conf.int = 0.95, digits = 4))

# --- BLOCO 2 ---

vars_bloco2 <- "CONSULTAS_cat + GESTACAO_cat"

formula_str_2 <- paste("desfecho ~", vars_bloco2)

# --- EXECUTAR A ANÁLISE (GLM nos dados imputados) ---
# Executar o modelo final nos 5 bancos de dados imputados

modelos_bloco_2 <- with(
  imputacao_mice,
  glm(as.formula(formula_str_2), family = "binomial")
)

# --- COMBINAR OS RESULTADOS ("POOL") ---
# Combinar (pool) os resultados

pool_bloco_2 <- pool(modelos_bloco_2)
print(summary(pool_bloco_2, exponentiate = TRUE, conf.int = 0.95, digits = 3))

# --- BLOCO 3 ---

vars_bloco3 <- "APGAR1_cat + APGAR5_cat + SEXO_cat + PESO_cat + IDANOMAL_cat"

formula_str_3 <- paste("desfecho ~", vars_bloco3)

# --- EXECUTAR A ANÁLISE (GLM nos dados imputados) ---
# Executar o modelo final nos 5 bancos de dados imputados

modelos_bloco_3 <- with(
  imputacao_mice,
  glm(as.formula(formula_str_3), family = "binomial")
)

# --- COMBINAR OS RESULTADOS ("POOL") ---
# Combinar (pool) os resultados

pool_bloco_3 <- pool(modelos_bloco_3)
print(summary(pool_bloco_3, exponentiate = TRUE, conf.int = 0.95, digits = 3))

# --- BLOCO FINAL ---

vars_bloco_4 <- "ESTCIVMAE_cat + QTDFILMORT_cat + IDADEMAE_cat + IDADEPAI_cat +
SEXO_cat + PESO_cat + IDANOMAL_cat"

formula_str_4 <- paste("desfecho ~", vars_bloco_4)

# --- EXECUTAR A ANÁLISE (GLM nos dados imputados) ---
# Execute o modelo final nos 5 bancos de dados imputados

modelos_bloco_4 <- with(
  imputacao_mice,
  glm(as.formula(formula_str_4), family = "binomial")
)

# --- COMBINAR OS RESULTADOS ("POOL") ---
# Combine (pool) os resultados

```

```
pool_bloco_4 <- pool(modelos_bloco_4)
print(summary(pool_bloco_4, exponentiate = TRUE, conf.int = 0.95, digits = 3))

summary_final <- summary(pool_bloco_4, exponentiate = TRUE, conf.int = 0.95)
summary_df <- as.data.frame(summary_final)

summary_df <- summary_df %>%
  mutate(across(c(estimate, std.error, p.value, `2.5 %`, `97.5 %`), ~round(.x, y)))
```

4. LASSO

```
#####
## LASSO ###
#####

library(pacman)
pacman::p_load(glmnet, dplyr, mice)

#####
## CONFIGURAÇÕES ###
#####

# Extrair UM banco de dados completo do MICE
banco_lasso <- complete(imputacao_mice, 1)

# Definir o Desfecho (Y)
y_lasso <- as.numeric(as.character(banco_lasso$desfecho))

# Preparar os Preditores (X) - Dummificação
# Precisamos transformar todos os fatores (_cat) em colunas 0/1
# Usamos model.matrix() para isso

# Crie a fórmula com todas as variáveis preditoras
formula_lasso <- as.formula(
  ~ ESCMAE_cat + ESTCIVMAE_cat + RACACORMAE_cat + IDADEMAE_cat +
    IDADEPAI_cat + QTDFILVIVO_cat + QTDFILMORT_cat + CONSULTAS_cat +
    GESTACAO_cat + GRAVIDEZ_cat + PARTO_cat +
    APGAR1_cat + APGAR5_cat + SEXO_cat + PESO_cat + IDANOMAL_cat
)

# O '-1' remove o intercepto, criando dummies para todas as categorias
x_lasso <- model.matrix(formula_lasso, data = banco_lasso)[, -1]

# --- Rodar o LASSO com Cross-Validation ---
# 'cv.glmnet' encontra o lambda (penalidade) ideal
# alpha = 1 significa LASSO (alpha = 0 seria Ridge)
# family = "binomial" é para regressão logística

cv_lasso <- cv.glmnet(
  x_lasso,
  y_lasso,
  family = "binomial",
  alpha = 1,
  type.measure = "auc" # Podemos otimizar pela AUC
)

# Ver o gráfico de cross-validation
plot(cv_lasso)

# --- Extrair os Coeficientes (As Variáveis Seleccionadas) ---
# Usamos 'lambda.1se' - é o lambda mais parcimonioso (seletivo)
# que ainda está dentro de 1 erro padrão do lambda ideal (lambda.min)

coeficientes_lasso <- coef(cv_lasso, s = "lambda.1se")

print(coeficientes_lasso)

# Variáveis têm um coeficiente NÃO-ZERO
variaveis_seleccionadas_lasso <- rownames(coeficientes_lasso)[
  coeficientes_lasso[, 1] != 0
]

# Remover o "(Intercept)" da lista
variaveis_seleccionadas_lasso <- variaveis_seleccionadas_lasso[
```

```
    variaveis_selecionadas_lasso != "(Intercept)"  
]  
  
print("--- Variáveis selecionadas pelo LASSO: ---")  
print(variaveis_selecionadas_lasso)
```