PAULA ANDREGHETTO BRACCO

Randomização Mendeliana: um método para estimação de efeitos causais utilizando variantes genéticas como variáveis instrumentais

Trabalho de conclusão de curso submetido como requisito parcial para a obtenção do grau de Bacharelado em Estatística

Orientador

Prof. Álvaro Vigo

Porto Alegre

2016

CIP - Catalogação na Publicação

Bracco, Paula Andreghetto
Randomização Mendeliana: um método para estimação de efeitos causais utilizando variantes genéticas como variáveis instrumentais / Paula Andreghetto Bracco. -- 2016.
61 f.

Orientador: Álvaro Vigo.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática, Curso de Estatística, Porto Alegre, BR-RS, 2016.

1. Randomização Mendeliana. 2. Causalidade. 3. Variáveis Instrumentais. I. Vigo, Álvaro, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Instituto de Matemática e Estatística

Departamento de Estatística

Randomização Mendeliana: um método para estimação de efeitos causais utilizando variantes genéticas como variáveis instrumentais

Paula Andreghetto Bracco

Banca examinadora:

Profª. Drª. Sídia Maria Callegari Jacques

UFRGS

Prof. Álvaro Vigo

UFRGS

AGRADECIMENTOS

À minha família por toda paciência e apoio. À minha mãe, Vera, que foi e sempre será minha maior incentivadora e aquela que acredita em mim acima de tudo.

Ao meu noivo, Bernardo, que em muitos momentos precisou ser compreensivo. Sem a sua ajuda no dia-a-dia com certeza terminar esse trabalho, e esse curso, teria sido muito mais desgastante.

Aos meus amigos, que são a melhor fonte de diversão e alívio para mente. Ao Gabriel e à Daiane, meus colegas e companheiros de curso que viraram amigos muito especiais também fora dos limites acadêmicos. Aos meus companheiros de trabalho e amigos do ELSA-Brasil, que me proporcionam crescimento pessoal, profissional e acadêmico.

Aos professores que me acompanharam durante a graduação, cada um contribuindo com o seu conhecimento para me tornar uma profissional mais completa, todos sempre dispostos e com vontade de ensinar e ajudar nos mais diversos problemas.

Em especial, agradeço ao meu orientador, Prof. Álvaro Vigo. Desde que solicitei conselho e ajuda sobre o caminho que eu gostaria de seguir na estatística, ainda no terceiro semestre, ele foi um verdadeiro conselheiro, professor e amigo. Esse trabalho foi um desafio e com certeza não teria sido possível concluí-lo sem a sua orientação, paciência e empenho.

“Agir, eis a inteligência verdadeira.

Serei o que quiser, mas tenho que querer o que for.

O êxito está em ter êxito, e não em ter condições de êxito.

Condições de palácio tem qualquer terra larga,

Mas onde estará o palácio se não o fizerem ali?”

Livro do Desassossego – Fernando Pessoa

APRESENTAÇÃO

O presente trabalho consiste em uma introdução à técnica de Randomização Mendeliana (RM). O assunto é extenso e bastante complexo, sendo impossível esgotar todas as suas características, aplicações e aspectos metodológicos nesse momento.

A Randomização Mendeliana é uma abordagem que tem sido muito utilizada para investigar causalidade em epidemiologia, com extensões para aplicações em epidemiologia molecular, sistemas biológicos, farmacogenética, etc. O trabalho apresenta de forma introdutória os principais conceitos no contexto de estimação de efeitos causais em estudos clínicos e epidemiológicos. Considerando a grande disponibilidade de dados de informações geradas em diversos consórcios *Genome-Wide Association Study* (GWAS), a atenção principal é dedicada ao contexto da aplicação da RM em dados sumarizados. Os achados destes estudos podem ser úteis para elaboração de novas hipóteses de pesquisa ou para o estudo de causalidade entre fatores de risco e desfechos de saúde.

Nesse trabalho, são introduzidos inicialmente conceitos genéticos, de causalidade e do uso de variáveis instrumentais para contextualização dos preceitos do método de RM. Na sequência, são descritas as principais abordagens, limitações e metodologias estatísticas para o uso da RM. No final, é exposto um exemplo de aplicação prática, com uma replicação de um estudo já publicado.

RESUMO

Atualmente, em estudos epidemiológicos, um grande desafio é a distinção entre associação e causalidade, e os ensaios clínicos randomizados (ECR) são considerados padrão-ouro para essa investigação. A aplicação de ECR, no entanto, é cara, longa e em muitas situações não é factível pela violação de preceitos éticos decorrentes da deliberada exposição a fatores de risco nocivos.

Nesse contexto, a randomização mendeliana (RM) é uma abordagem alternativa para estimar a relação causal entre exposições biológicas modificáveis ou fatores biológicos intermediários e um desfecho clínico de interesse, utilizando variantes genéticas (*Single Nucleotide Polymorphism* - SNPs) como variáveis instrumentais. A técnica de variáveis instrumentais (VI) é uma das abordagens possíveis para estimação de efeito causal, mesmo sem o conhecimento de todos os possíveis confundidores presentes na associação entre exposição e desfecho

Nos estudos epidemiológicos de RM, os SNPs são utilizados como instrumentos e a estrutura do delineamento é semelhante a um ECR. A randomização ocorre durante a segregação genética e os grupos de exposição são definidos pela presença do alelo mais frequente ou do alelo variante do SNP, gerando grupos que não diferem entre si quanto aos fatores de confundimento

Apesar do aumento no número de estudos de associações genéticas (GWAS) entre SNPs e desfechos clínicos e/ou exposições biológicas, muitas vezes pode ser difícil obter bases de dados no nível dos indivíduos com informações suficientes. A abordagem de RM em utilizando dados sumarizados permite combinar resultados já publicados em estudos anteriores, tornando-se uma alternativa relevante para investigação de causalidade.

A interpretação dos resultados de aplicação da RM, no entanto, ainda deve ser realizada com cuidado, considerando as possíveis limitações do método.

O objetivo desse trabalho é introduzir os conceitos e princípios da randomização mendeliana, trazer exemplos e aplicações em estudos epidemiológicos e explicar os métodos de estimação de causalidade pelo uso de variáveis instrumentais.

Palavras-chave: Randomização Mendeliana, Causalidade, Variáveis Instrumentais

Abstract

A major challenge in epidemiological studies is the distinction between association and causality, and randomized controlled trials (RCTs) are considered the gold standard for this kind of investigation. The application of ECR, however, is expensive, lengthy and in many situations infeasible by the lack of ethical precepts in deliberately exposing individuals to harmful risk factors.

In this context, Mendelian Randomization (MR) is an alternative approach to estimate the causal relationship between modifiable biological exposures or intermediate biological factors and a clinical outcome of interest, using genetic variants (Single Nucleotide Polymorphism - SNPs) as instrumental variables. The technique of instrumental variables (VI) is one of the possible approaches for estimation of causal effect, even without the knowledge of all possible confounders present in the association between exposure and outcome.

In epidemiological studies of MR, SNPs are used as instruments and the study structure is similar to an ECR. Randomization occurs during genetic segregation and exposure groups are defined by the presence of the most frequent allele or the variant allele of the SNP, generating groups that do not differ in respect to confounding factors.

Despite the increase in genetic association studies (GWAS) between SNPs and clinical outcomes and/or biological exposures, it can often be difficult to obtain data at individual levels with sufficient information. The MR approach using summarized data allows the use of summarized results already published in previous studies, and it is considered a relevant alternative for investigating causality.

The interpretation of the results of MR application, however, should still be performed with caution, considering the possible limitations of the method.

The objective of this work is to introduce the concept and principles of Mendelian randomization, to bring examples and applications in epidemiological studies and to explain causal estimation methods by the use of instrumental variables.

Key-Words: Mendelian Randomization, Causality, Instrumental Variables

Sumário

1. Introdução

Em estudos clínicos e epidemiológicos que visam determinar a etiologia de uma doença, um grande desafio é a distinção entre associação e causalidade. Em outras palavras, identificar se a estimativa da associação entre uma exposição e o desfecho de fato representa um efeito causal (Burgess and Thompson, 2015). Neste contexto, os ensaios clínicos randomizados são considerados padrão-ouro para investigar hipóteses científicas em pesquisa clínica e epidemiológica. No entanto, eles costumam ser caros e longos, ou não são factíveis pela impossibilidade (por razões éticas, por exemplo) de alocar os indivíduos ao acaso aos grupos que representam as intervenções. Ainda, muitas vezes os resultados não podem ser generalizados para populações fora dos critérios de inclusão do estudo (Sanson-Fisher et al., 2007).

Assim, estudos observacionais, como estudos longitudinais de coorte, caso controle, caso-coorte ou caso-controle aninhado em uma coorte, frequentemente são usados para avaliar uma hipótese de causalidade. No entanto, estão sujeitos a confundimento, causalidade reversa e diversos vieses que, muitas vezes, impossibilitam verificar se uma exposição de fato causa determinada doença. Mesmo estudos observacionais robustos que satisfazem os critérios empíricos descritos por Hill (Tabela 1) para inferir casualidade (Hill, 1965) são suscetíveis a esses vieses, muitas vezes ocasionados também por confundimento não observável (*unmeasured confounding*).

Tabela 1. Critérios de causalidade entre doença e fator ambiental (HILL, 1965)

Força da associação
Reprodutibilidade
Especificidade
Temporalidade
Gradiente biológico
Plausibilidade biológica
Coerência
Evidência Experimental
Analogia

Existem diversos exemplos de associações entre fatores de risco e doenças, estimados por meio de estudos observacionais que depois se mostram sem efeito causal. Um

exemplo clássico são os estudos de reposição hormonal pós-menopausa. Estudos observacionais iniciais sugeriram benefícios dessa terapia em diversas doenças, entre elas as doenças cardiovasculares, inferindo associação dessa terapia com diminuição de LDL (*low density lipids*) e aumento de HDL (*high density lipids*) (Barrett-Connor et al., 1997). Em 2002, no entanto, o ensaio clínico randomizado de grande porte '*Women's Health Initiative*' publicou seus resultados após anos de acompanhamento de mulheres em intervenção de tratamento de reposição hormonal, concluindo que os riscos dessa terapia excediam os benefícios (Rossouw et al., 2002) e que o risco de doenças cardiovasculares aumentava 29% quando comparado com o placebo. Em 2012, após diversos estudos terem sido realizados sobre esse assunto, Davey (2012) revelou que os riscos de doença cardiovascular dependiam da idade do inicio do tratamento de reposição hormonal, da dose, da rota de administração, e concluiu que o tratamento de reposição hormonal tem vantagens, mas não é recomendado para mulheres acima de 60 anos (Davey, 2012). As conclusões iniciais anteriores a esse estudo guiaram por muito tempo um protocolo de tratamento para mulheres na menopausa, o qual posteriormente se mostrou inadequado, devido principalmente aos confundidores que não puderam ser devidamente controlados, levando à estimativas viesadas de magnitude das associações (Wannmacher and Lubianca, 2004).

A randomização mendeliana é uma abordagem alternativa para estimar a relação causal entre exposições biológicas modificáveis ou fatores biológicos intermediários e um desfecho clínico de interesse, utilizando variantes genéticas como variáveis instrumentais (Evans and David Smith, 2015).

Embora não tenha sido chamado de randomização mendeliana, a primeira descrição deste conceito foi apresentada em 1986 por Martijn Katan, em cujo trabalho foi discutido se níveis baixos de colesterol sérico realmente têm relação causal com câncer, ou se representava um caso de causalidade reversa ou confundimento devido às características de dieta ou de outros fatores (Katan, 1986, 2004). Posteriormente o método foi denominado de "Randomização Mendeliana" e aplicado no contexto de transplante alogênico de medula óssea, visando comparar a sobrevivência de pacientes HLA compatíveis (*HLA-Human Leukocyte Antigen*) com pacientes não compatíveis, sem a necessidade de realizar um ensaio clínico randomizado (Davey Smith and Ebrahim, 2003; Gray and Wheatley, 1991). Do ponto de vista estatístico, o método é uma aplicação da técnica de variáveis instrumentais, em que as variantes genéticas (diferentes versões de um mesmo gene, herdadas, presentes em todas as células e não provenientes de mutação) atuam como instrumentos da exposição de interesse (Burgess and Thompson, 2015; Didelez and Sheehan, 2007). Além da epidemiologia, áreas como biologia molecular, farmacogenética, entre outras, já estão fazendo uso desse desenho de estudo.

O objetivo desse trabalho é introduzir os conceitos e princípios da randomização mendeliana, trazer exemplos e aplicações de estudos epidemiológicos e explicar os métodos de estimação de causalidade pelo uso de variáveis instrumentais.

2. Conceitos Básicos

2.1. Causalidade e inferência causal

De maneira geral, no contexto estatístico pressupõe-se que uma estimativa de correlação ou de associação não implica necessariamente causalidade. Interpretações equivocadas dessas medidas ocorrem frequentemente, especialmente em estudos epidemiológicos delineados de forma que não possibilitem a realização de inferência causal (Burgess and Thompson, 2015).

Em 1985, P.W. Holland apresentou uma discussão sobre a estatística, causalidade e inferência causal, descrevendo o modelo de inferência causal proposto por D.B. Rubin. O autor revisou conceitos de inferência causal apresentados por outros filósofos, incluindo Aristóteles, Hume, Mill e Suppes, bem como por estatísticos tais Kempthorne, Cox, Fisher e Neyman. Finalmente, discutiu os conceitos de inferência causal em ciências como Medicina, Economia e Ciências Sociais (Holland et al., 1985).

Os conceitos de causalidade e inferência causal e as abordagens filosóficas comumente utilizadas são demasiadamente extensos e complexos para que se pudesse abordá-los com profundidade neste trabalho. Neste sentido, esta seção apresenta conceitos básicos necessários para discutir o método de randomização mendeliana. Para o leitor interessado em aprofundar seus conhecimentos sobre o assunto, sugere-se consultar, por exemplo, o livro Epidemiologia Moderna, organizado por Rothman, Greenland e Lash (Rothman et al., 2008). Nesse livro, as principais abordagens filosóficas para o estudo da causalidade e da inferência causal são apresentadas no Capítulo 2 (modelo de causa suficiente componente), Capítulo 4 (modelo de causalidade potencial-desfecho, ou contrafactual) e Capítulo 12 (modelos causais gráficos, ou diagramas causais).

O modelo contrafactual de estudos de causalidade contrasta o desfecho ocorrido sob condições específicas, com o desfecho alternativo que ocorreria em condições alternativas. Por exemplo, o efeito causal de um tratamento T em um determinado desfecho pode ser definido como a proporção de desfecho ocorrido com o uso de T menos a proporção de desfecho ocorrido sem o uso de T, dado que todas as demais condições permanecessem constantes. Essa definição sozinha, no entanto, é insuficiente para definir causalidade, e deve ser considerara como uma informação adicional que, quando utilizada em conjunto com as demais abordagens de inferência causal, permite fortalecer a distinção entre uma associação e um efeito causal (Parascandola, 2001).

Na abordagem do modelo de causa suficiente, a causa de uma doença específica pode ser definida como um evento, condição ou característica que além de preceder o início

da doença em questão também é necessário para a sua ocorrência, dado que as demais condições se mantêm constantes. Dentro desse raciocínio, pode ocorrer que um determinado evento, condição ou característica não seja suficiente, isoladamente, para o desenvolvimento do desfecho em questão, não sendo considerado um mecanismo causal completo e sim um componente do mesmo. Denomina-se 'causa suficiente' o conjunto de mecanismos causais completos, definidos como o conjunto de condições e eventos mínimos necessários para o aparecimento inevitável do desfecho, sendo possível um determinado componente participar de diversos mecanismos causais (Figura 1) (Rothman and Greenland, 2005). Essa definição é muito utilizada para inferência causal de doenças infecciosas, com vírus e bactérias representando os principais componentes do mecanismo causal.

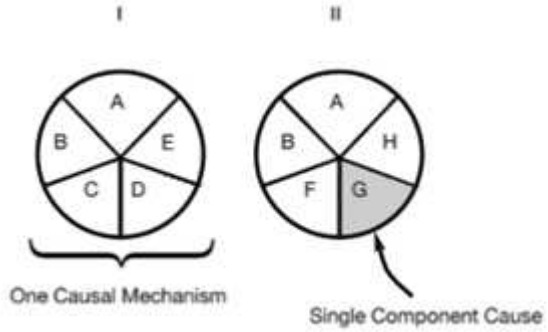

Figura 1. I e II representam duas causas suficientes, isto é, dois mecanismos causais completos compostos com componentes causais diferentes.
Fonte: Adaptado de *Rothman and Greeland, 2005*.

No entanto, definições precisas no contexto epidemiológico de doenças crônicas são complexas, especialmente porque a causa muitas vezes é probabilística e não determinística. O fumo, por exemplo, pode resultar em câncer de pulmão, porém podem existir indivíduos que fumam e nunca desenvolvem a doença, bem como indivíduos acometidos por essa neoplasia que nunca experimentaram tabaco (Burgess and Thompson, 2015).

Hernán (2004) revisou e discutiu uma definição de efeito causal na pesquisa epidemiológica, exemplificando a estimação de efeito causal no contexto de resposta dicotômica. A discussão é estendida para aspectos como causalidade versus associação, o uso da randomização e suas limitações e presença de variabilidade amostral (Hernan, 2004).

Modelos gráficos baseados nos gráficos diretos acíclicos (*DAG – Directed Acyclic Graphs*) muitas vezes são úteis para a representação das relações causais. Os gráficos são compostos de 'nodos' que representam as variáveis em questão (por exemplo, um nodo 'X' representando uma variável de exposição, um nodo 'C' representando uma variável confundidora e um nodo 'Y' representando a variável de desfecho) e de flechas que ligam essas variáveis. Os DAGs não permitem a ocorrência de ciclos, ou seja, que uma sequência direta de flechas ligue um nodo a ele mesmo. Uma flecha da variável X para a variável Y (X → Y) indica que existe um efeito causal de X em Y (Burgess and Thompson, 2015; Didelez and Sheehan, 2007)

O modelo de DAG mais simples e que servirá de base para os métodos abordados nesse trabalho é representado na Figura 2. Outros modelos mais complexos são descritos na literatura (Greenland et al., 1999).

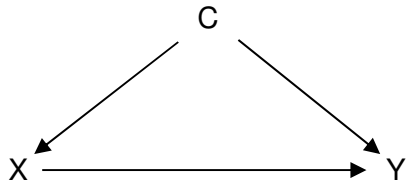

Figura 2. Modelo DAG representando efeito causal de X em Y e os efeitos do confundidor C em X e em Y.

Definições formais dos conceitos de diagramas causais aplicados à pesquisa epidemiológica foram apresentados por Greenland et al (1999), que discutiram aspectos como causalidade e associação, e confundimento. Greenland e Brumback (2002) apresentaram uma revisão sobre as quatro principais abordagens para modelagem de efeitos causais: diagramas causais, modelo contrafactual, modelo causa suficiente e equações estruturais. Os diagramas causais podem ser úteis para explicitar de forma clara as suposições conceituais envolvidas na análise da causalidade, ao passo que os modelos contrafactual e de equações estruturais podem explicitar as relações funcionais de forma quantitativa. O modelo de causa suficiente difere dos demais por apresentar as premissas conceituais do mecanismo causal de forma mais detalhada. Os autores fazem relações entre os modelos, as representações gráficas e algébricas para o modelo de causalidade, explicitando também o uso de variáveis instrumentais para estimar os efeitos causais (Greenland and Brumback, 2002).

Hernán e Robins (2006) também exploraram o uso de variáveis instrumentais para estimar efeitos causais, mesmo na presença de confundimento não mensurável. Eles

descreveram as conexões entre quatro modelos de causalidade (modelo contrafactual, DAGs, modelo não paramétrico de equações estruturais e modelo linear de equações estruturais), e formalizaram uma abordagem unificada usando variáveis instrumentais (Hernán and Robins, 2006).

Existem ainda outras abordagens para estimação de efeitos causais em estudo epidemiológicos, como, por exemplo, os modelos estruturais marginais, porém também estão fora do escopo deste trabalho (Robins et al., 2000)

2.2. Conceitos Genéticos

A informação genética dos humanos está contida em 23 pares cromossomos, cuja principal função é a de portar os genes, unidades funcionais da hereditariedade. Um gene é um segmento de DNA que serve como molde para uma molécula de RNA funcionalmente importante. As informações biológicas contidas nos genes devem ser copiadas com precisão para que sejam transmitidas às células-filhas e, consequentemente, às próximas gerações. O DNA é uma dupla hélice formada por duas fitas complementares de nucleotídeos que são mantidas unidas por pontes de hidrogênio entre os pares de base G-C e A-T (Alberts, 2004). A definição de gene tem mudado e evoluído conforme suas propriedades vão sendo conhecidas, no entanto, para essa monografia, um gene será considerado um determinante, ou codeterminante, de uma característica que é herdada segundo as leis de Mendel (King et al., 2013).

Com exceção das células germinativas (óvulos e espermatozoides), em geral as células possuem duas cópias de cada cromossomo, uma herdada da mãe e a outra do pai, sendo denominados cromossomos homólogos. Os cromossomos homólogos geralmente são bastante similares entre si, no entanto, diferentes versões de um mesmo gene podem ocorrer em cada cromossomo, as quais são denominadas alelos. Um indivíduo portador de dois alelos diferentes, um em cada cromossomo, é denominado heterozigoto, sendo definido como homozigoto caso porte dois alelos iguais (Alberts, 2004).

Entre os seres humanos pode haver variabilidade na sequência do DNA, geralmente na forma de polimorfismos de nucleotídeo único (SNP). Um SNP é uma variação de um nucleotídeo em um loco específico de uma sequência de DNA, sendo que essa variação pode ocorrer em genes ou em regiões não codificadoras. Para ser classificada como SNP a variação deve ocorrer no mínimo em 1% dos indivíduos da população. Nem todos SNPs acarretam problemas, mas certos SNPS estão associados com desordens e doenças, tais como hipertensão, obesidade e câncer (SNP | Learn Science at Scitable). Muitos sítios do

DNA mapeados no genoma humano são polimórficos, significando que existe uma probabilidade razoável de que os genomas de dois indivíduos irão diferir nesses sítios, o que os tornam úteis para análises que investigam a associação de características específicas (fenótipos) com sequências de DNA específicas (Alberts, 2004).

2.2.1.Segunda Lei de Mendel

Quando a células germinativas são criadas pelo processo de meiose, ocorre uma redução do número de cromossomos, isso é, nas células reprodutivas (gametas) existe apenas um exemplar de cada um dos 23 cromossomos. Conhecida como lei da segregação independente, a 2º lei de Mendel afirma que a escolha de qual cromossomo homólogo do par formará o gameta é aleatória e independente entre os diferentes pares de cromossomos (Glennan, 1996).

Como a segregação e a transferência aos gametas dos cromossomos homólogos, e consequentemente dos alelos e dos possíveis polimorfismos, ocorre de forma aleatória e independente, variações genéticas (como os SNPs) associadas com exposições modificáveis e/ou com desfechos de saúde são independentes de outras características determinadas pelos demais genes e/ou variações genéticas presentes, o que evita possíveis problemas de confundimento na medida de efeito dessa associação. Além disso, como a segregação ocorre na formação dos gametas, tal associação também não está sujeita a viés de causalidade reversa.

2.2.2. Interação Gene-Ambiente

Quando se estima separadamente a contribuição individual dos genes e do ambiente, a proporção da doença explicada pelo efeito conjunto da interação gene-ambiente é ignorada (Hunter, 2005). Algumas interações entre gene e ambiente puderam ser observadas mesmo sem o uso de análises moleculares. Um exemplo conhecido é o efeito da exposição ao sol na incidência de câncer de pele, sendo que indivíduos de pele clara são mais susceptíveis a essa exposição (Hans-Olov, 2008). Com a conclusão do Projeto Genoma Humano, que identificou e publicou a sequência completa do DNA humano com todas as cerca de três bilhões de bases de nucleotídeos, tornou-se possível a investigação de diferenças entre indivíduos no nível molecular, expandindo o território para o estudo da interação gene-ambiente (Collins et a., 2003).

Estudos convencionais de epidemiologia genética investigam a base genética de determinada característica ou a função de determinado gene, utilizando os SNPs como marcadores de variação genética. Mas também é possível explorar o conceito de segregação independente, utilizando os genes como uma ferramenta para reduzir o confundimendo existente nas associações entre expositores modificáveis, muitas vezes provenientes do ambiente e/ou de hábitos de vida, e doenças de interesse. Esta abordagem, denominada Randomização Mendeliana (Davey Smith and Ebrahim, 2003), tem como princípio o de que se um determinado SNP está relacionado a efeitos biológicos resultantes de exposições modificáveis e também altera o risco a determinada doença, então a exposição está relacionada de forma causal à doença. Assim, SNPs com uma função biológica bem definida podem ser utilizados para estudar o efeito de exposições suspeitas no aumento do risco de doenças (Davey Smith and Ebrahim, 2003).

2.3. Uso de variantes genéticas como variáveis instrumentais

A técnica de variáveis instrumentais (VI) é uma das abordagens possíveis para estimação de efeito causal, mesmo sem o conhecimento de todos os possíveis confundidores presentes na associação entre exposição e desfecho (Burgess and Thompson, 2015)

Para que uma variante genética (SNP) possa ser utilizada como variável instrumental, deve satisfazer determinadas condições representadas no diagrama causal exposto na Figura 3, na qual a letra ‘G’ representa a variante genética (Burgess and Thompson, 2015 ; Haycock et al., 2016) :

I. Estar associada a variável expositora ‘X’,
II. Não estar associada com algum confundidor da associação exposição-desfecho, representado pela letra ‘C’;
III. Não afetar o desfecho, exceto pelo efeito da exposição;

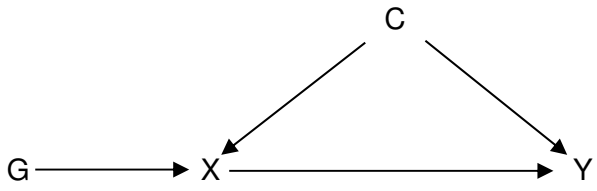

Figura 3. Modelo DAG representando o uso de uma variante genética G como variável instrumental. A exposição X possui efeito causal no desfecho Y; a variante G está associada com X, é independente dos possíveis confundidores C e não está associada com Y, exceto via X.

As possíveis variações de nucleotídeos em um SNP são denominadas alelos e a maioria dos SNPs é dialélica, com um alelo sendo mais comum (*wild-type ou major*) e o outro, menos comum (*minor*), considerado o alelo variante. Para cada indivíduo o genótipo de um SNP pode ser representado matematicamente como uma variável aleatória que expressa o número de alelos variantes (0, 1 ou 2, uma vez que devido as duas cópias de DNA presente nas células humanas, um indivíduo pode no máximo portar 2 alelos incomuns). As medidas de associação do SNP com os expositores podem ser interpretadas como a mudança no fator de risco a cada alelo variante adicional (Burgess and Thompson, 2015; Burgess et al., 2015a) Entre as principais razões para a violação das premissas para a utilização de variantes genéticas como variáveis instrumentais estão mecanismos biológicos, genéticos ou populacionais. A Tabela 2 resume os aspectos do mecanismo que causa a violação, seus efeitos e formas de controle (Burgess and Thompson, 2015).

Tabela 2. Principais possíveis violações das premissas para uso de variáveis genéticas como variáveis instrumentais

Mecanismo	Efeito	Controle
Biológico		
Pleiotropia	A variante genética está associada com diversos fatores de risco (isto é, um mesmo par de alelos está associado a diversas características), levando a violação das condições II e III.	Utilizar variáveis genéticas localizadas em genes com função biológica bem conhecida
Canalização	O indivíduo desenvolve mecanismos para compensar o efeito causado pela variação genética. Torna difícil verificar a relação entre variação genética e exposição de interesse.	Não é uma violação propriamente dita, mas sim uma consequência indesejável. Podem-se considerar as alterações no fator de risco devido a canalização provenientes de um efeito causal da variante genética.
Genético		
Desequilíbrio de Ligação (LD)	Variantes genéticas muito próximas em um mesmo cromossomo tendem a ser herdadas em conjunto, não respeitando a lei da segregação independente de Mendel. Variantes correlacionadas com a variante genética usada na análise podem ter efeitos sobre fatores de risco competitivos, podendo ocasionar a violação das condições II e III.	A variável genética utilizada como variável instrumental não precisa ser obrigatoriamente a variante causal, devendo apenas estar correlacionada a ela.
Modificação de efeito	Interação estatística entre o efeito da exposição e uma covariável. Assim, o efeito causal varia de acordo com o estrato da covariável em questão.	Para interpretação do efeito causal deve-se considerar a interação e como o mesmo se comporta em cada estrato da covariável.
Populacional		
Estratificação	A distribuição da variante genética e da exposição são diferentes entre subpopulacões, por exemplo, grupos de origem étnica distinta.	Restringir os estudos a populações de mesma etnia/ancestralidade

Fonte: Burgess and Thompson (2015)

A validade de uma variante genética como VI só pode ser analisada de forma empírica, e não conclusiva, partindo da análise da associação dessa variante com confundidores conhecidos e avaliando se a associação entre a variante e o desfecho é atenuada de forma significativa quando ajustada pela exposição. Sendo assim, a base para justificar a escolha de determinada variante genética deve sempre ser o conhecimento biológico. Os critérios de causalidade de Bradford Hill também podem ser utilizados para julgar a plausibilidade de uma variante genética como VI (Burgess and Thompson, 2015). Sucintamente, os autores descrevem estes critérios da seguinte forma:

- ***Força:*** Uma associação fraca entre a variante genética e o desfecho pode ser explicada por um pequeno desequilíbrio em uma covariável associada à variante genética. Violações leves nas suposições da VI são menos prováveis de serem identificadas por meio de testes de associações entre a variante genética e as covariáveis.
- ***Consistência:*** A relação de casualidade é mais plausível se múltiplas variantes genéticas associadas com a mesma exposição também estão associadas com o desfecho na mesma direção e aproximadamente mesma magnitude, especialmente se as variantes estão localizadas em diferentes regiões do gene e/ou se os mecanismos da associação com o desfecho são distintos.
- ***Gradiente biológico:*** A relação de causalidade é mais plausível se as associações genéticas com o desfecho e com a exposição são proporcionais para cada variante.
- ***Especificidade:*** A relação da causalidade é mais plausível se as variantes genéticas estão associadas com um específico fator de risco e desfecho, e se não existirem associações com muitas covariáveis e outros desfechos.
- ***Plausibilidade:*** Se a função da variante genética é conhecida, então a causalidade é mais plausível se o mecanismo pelo qual a variante genética atua está relacionada com a exposição com credibilidade e de forma específica.
- ***Coerência:*** Se foi realizada uma intervenção na exposição, associações observadas em um contexto experimental com desfechos (ou covariáveis) intermediários também devem estar presentes no contexto genético.

Burguess e Thompson (2015) formalizam matematicamente a definição de VI como uma variável aleatória. Utilizando o contexto do modelo DAG especificado na Figura 3, o desfecho Y é uma função de uma exposição mensurável X e de um confundidor não mensurável C (nesta representação, pressupõe-se que os fatores de confusão podem ser resumidos em uma única variável aleatória C). Ainda, a exposição X pode ser expressa como

uma função do confundidor C e de uma variante genética G (aqui, G representa uma ou mais variantes). As condições I, II e III necessárias para que a variante G possa ser utilizada como VI, descritas anteriormente, podem ser formalizadas por meio das variáveis aleatórias, como segue:

I. G não é independente de X ($G \not\perp\!\!\!\perp X$);
II. G é independente de C ($G \perp\!\!\!\perp C$); e,
III. Condicional em X e C, G é independente de Y ($G \perp\!\!\!\perp Y|X,C$).

As suposições I, II e III implicam que a distribuição conjunta de Y, X, C e G pode ser fatorada como

$$p(y, x, c, g) = p(y|c, x)\, p(x|c, g)\, p(c)\, p(g),$$

que corresponde ao modelo DAG da Figura 3. Burguess e Thompson (2015) contextualizaram esta definição formal de VI como variável aleatória para a abordagem do modelo de causalidade, porém a definição completa está além dos objetivos propostos neste trabalho.

Se as suposições I, II e III estão atendidas, então a variante genética G pode ser utilizada como VI para testar a hipótese de causalidade entre a exposição X e o desfecho Y. Uma associação não nula entre a variante G e a exposição X, e entre a variante G e o desfecho Y, sugere a presença de relação causal entre X e Y. Se a associação entre a variante G e o desfecho Y for nula, não há evidencia de que existe correlação linear entre G e Y. Porém, como correlação linear zero não implica necessariamente em independência, pode não ser correto afirmar que não existe relação causal entre a exposição e o desfecho, especialmente quando são analisados modelos biológicos plausíveis (Burgess and Thompson, 2015).

Sendo assim, em vez de somente testar a existência de relação causal, é vantajoso estimar a magnitude do efeito. Quando são utilizadas diversas variantes genéticas pode haver um aumento considerável de poder. Por outro lado, mesmo se a estimativa do efeito causal não for significativa, podem ser usados os limites do intervalo de confiança para julgar a relevância clínica do efeito causal.

Duas suposições adicionais são exigidas para estimar o efeito causal (Burgess and Thompson, 2015). A primeira especifica que o desfecho potencial para cada indivíduo não deve ser influenciado pela forma como a exposição ocorre, e também não pode ser influenciado pelas covariáveis do modelo relativas aos outros indivíduos. Esta suposição geralmente não é plausível no contexto da randomização mendeliana e, portanto, a estimativa de efeito não pode ser interpretada, de forma simplista, como o resultado esperado da

intervenção em uma exposição ou fator de risco. A segunda, chamada de monotonicidade forte (*strong monotonicity*), postula que a variação da VI deve modificar a exposição para ao menos um indivíduo na população e que qualquer modificação na exposição devido à variação na VI deve estar na mesma direção (aumento ou diminuição) para todos os indivíduos. A monotonicidade forte tende a ser atendida para a maioria das aplicações de Randomização Mendeliana em que existe plausibilidade biológica (Burgess and Thompson, 2015).

3. Randomização Mendeliana

A randomização mendeliana (RM) é uma nova forma de síntese de evidência e inferência causal que vem ganhando importância no campo da epidemiologia observacional (Haycock et al., 2016). Atualmente, genótipos podem ser analisados com alta sensibilidade e acurácia, e estudos com SNPs podem revelar efeitos de exposições crônicas e agudas. No entanto, diferente dos estudos GWAS em que o objetivo é demonstrar a associação entre uma variante genética e um desfecho, na abordagem da RM o foco é fornecer evidencias a favor ou contra a relação causal entre uma exposição modificável e um desfecho de saúde de interesse, além de estimar a magnitude dessa relação (Evans and David Smith, 2015).

Estudos com RM têm sido realizados para estimar o efeito causal de diversos expositores biológicos em doenças crônicas, geralmente obtendo-se resultados concordantes com aqueles baseados em ensaios clínicos randomizados (ECR) (C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC), 2011; Palmer et al., 2013; Voight et al., 2012).

Para facilitar o entendimento dos princípios da RM é conveniente fazer uma analogia com os ECR (Figura 4). Por exemplo, em estudos de ECR os participantes são alocados aleatoriamente para um de dois grupos, tratamento ou controle. A aleatorização visa criar grupos comparáveis e, assim, eliminar (tanto quanto possível) vieses de confundimento, seja por características observáveis ou não observáveis. A estrutura do delineamento na RM é semelhante, e a randomização ocorre durante a segregação genética. Assim, os grupos de exposição são definidos pela presença do alelo mais frequente ou do alelo variante do SNP, gerando grupos que não diferem entre si quanto aos fatores de confundimento (Davey Smith and Hemani, 2014).

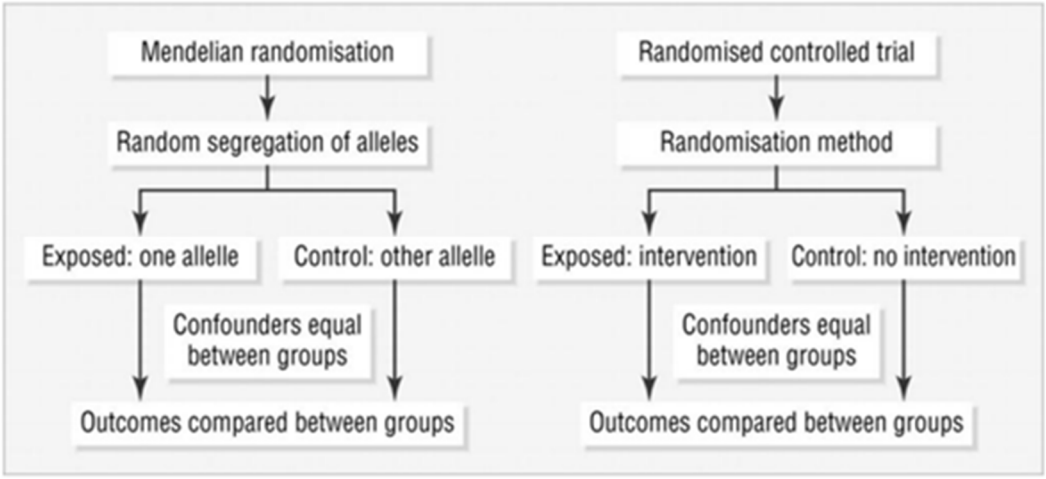

Figura 4. Comparação entre os delineamentos RM e ECR .
Fonte: Adaptado de (Davey and Smith, 2005).

Além disso, como a segregação genética ocorre previamente às exposições ambientais, por exemplo, não há possibilidade de a associação entre a variante genética e o desfecho ocorrerem devido à causalidade reversa.

A RM é considerada uma aplicação da análise de variáveis instrumentais. Nos estudos epidemiológicos os SNPs são utilizados como instrumentos e as premissas descritas no Capítulo 2 necessitam ser verdadeiras. O passo mais importante em um estudo de RM é definir o instrumento genético, o qual deve estar associado de forma robusta com a exposição de interesse. Para tanto, as duas principais abordagens são utilizar SNPs presentes em genes, que tenham efeito biológico bem conhecido, ou procurar SNPs a partir de resultados de GWAS disponíveis. A segunda abordagem deve ser utilizada com mais cautela, uma vez que muitos SNPs ainda não possuem um embasamento biológico forte e podem ser mais susceptíveis a violações dos requisitos para serem usados como variáveis instrumentais (Haycock et al., 2016).

3.1. Principais abordagens utilizadas na RM

Qualquer abordagem que utilize informações genéticas com objetivo de inferir causalidade entre uma exposição de interesse (X) e um desfecho (Y) pode ser considerada uma aplicação da RM. Esta seção descreve sucintamente as principais estratégias de aplicação da RM para estudos de causalidade. Como ilustração, a Tabela 3 resume informações de alguns estudos, descrevendo a abordagem utilizada na RM e o assunto investigado.

Associação Gene-Exposição e Gene-Desfecho

É o estudo com desenho mais simples, no qual a existência de associação entre um gene e a exposição de interesse e entre esse mesmo gene e o desfecho implica uma relação causal exposição-desfecho. Essa abordagem não permite estimar a magnitude da associação causal (Haycock et al., 2016).

RM em uma amostra com dados individuais

Geralmente utiliza banco de dados provenientes de estudos GWAS, nos quais estão presentes informações no nível individual de um amplo rastreamento de SNPs, sobre nível/presença da exposição e nível/presença do desfecho. Nesses casos, o efeito causal pode ser estimado (Haycock et al., 2016).

Tabela 3. Exemplos de estudos que utilizaram a RM para estimação de efeito causal e respectiva abordagem.

Abordagem do Estudo	Descrição	Referência
Associação X-G e Y-G	SNP no gene LPA associado com a concentração de lipoproteína Lp(a) e com doença coronariana; Relação causal positiva entre concentração de Lp(a) e risco de doença coronariana.	Clarke et al., (2009)
RM em uma amostra de dados individuais	4 variantes genéticas estudadas em uma mesma amostra não encontrou relação causal entre proteína C reativa e depressão	Wium-Andersen et al., (2014)
RM em duas amostras de dados individuais	Primeira amostra: análise da associação entre SNPs no gene LIPG com HDL. Segunda amostra: análise de presença de associação dos SNPs identificados na primeira amostra com infarto do miocárdio. Não foi observada relação causal entre o aumento do HDL e menor risco de infarto.	Voight et al., (2012)
RM em duas amostras de dados sumarizados	Primeira amostra: Estudos com dados sumarizados de associação de SNPs de genes relacionados a síntese e metabolismo de vitamina D e níveis circulantes dessa vitamina. Segunda amostra: Estudos com dados sumarizados desses SNPs com risco de diabetes tipo 2. Não foi observada relação causar entre concentração de vitamina D e diabetes tipo 2.	Ye et al., (2015)
RM bidirecional	Variantes genéticas associadas ao IMC (Índice de Massa Corporal) e variantes associadas a atividade física. Foi observada relação causal entre aumento do IMC com redução da atividade física, mas não o contrário.	Evans and David Smith (2015)
RM em dois passos	Primeiro passo: Foi avaliada a associação entre SNPs relacionados com glicemia materna e metilação em um *locus* do gene LEP. Segundo passo: Foi avaliada a associação entre a metilação do gene LEP e a concentração de leptina no cordão umbilical. Foi observada relação casal entre níveis de glicemia materna e níveis de leptina fetais.	Allard et al., (2015)

RM utilizando duas amostras

A abordagem de duas amostras é uma extensão do procedimento de mínimos quadrados em dois estágios (*2SLS-Two Stages Least Squares*), a ser descrito na Seção 3.2, e amplia o potencial da RM como método para estimação de efeitos causais. Nesse contexto, as associações entre X-G e Y-G são estimadas em amostras separadas. A principal premissa a ser atendida é que as amostras não possuam indivíduos em comum e representem a mesma população, principalmente quanto à idade, sexo e etnia (Haycock et al., 2016).

O método pode ser utilizado tanto com informações individuais quanto com dados sumarizados, quando somente é possível acessar as estimativas de associação X-G e Y-G, não sendo conhecidos os dados de cada indivíduo do estudo. Neste caso, é importante assegurar que as associações X-G e Y-G considerem o mesmo alelo variante.

Considerando a crescente disponibilização de informações geradas por consórcios de GWAS de grande porte, dados provenientes de diferentes estudos podem ser utilizados, aumentando-se assim o tamanho amostral, de modo que o poder estatístico e a precisão das estimativas também aumentam.

RM bidirecional

Neste enfoque de análise, variáveis genéticas diferentes e independentes são utilizadas como variáveis instrumentais para cada fenótipo de interesse para investigar a direção da relação causal. Essa metodologia é utilizada especialmente em casos em que não se conhece a função biológica das variáveis genéticas utilizadas (Evans and David Smith, 2015).

RM em dois passos

Esta abordagem é utilizada para avaliar em que magnitude uma relação causal entre X e Y pode ser mediada por uma variável intermediária (Z). Foi desenvolvida para o uso em estudos epigenéticos, isto é, para analisar o quanto alterações químicas no DNA poderiam mediar relações importantes entre exposições e desfechos de saúde relevantes. Essas alterações, em especial metilações e demetilações, geralmente são responsáveis pela alteração da expressão gênica sem a presença de mutações e/ou variantes genéticas. No primeiro passo, variantes genéticas associadas à exposição X são utilizadas para avaliar a relação causal entre X e uma variável Z intermediária, possivelmente mediadora do efeito de X em Y. No segundo passo, instrumentos genéticos diferentes, associados a Z, são utilizados para avaliar o efeito causal de Z em Y. A exposição, o mediador e o desfecho não podem ser medidos no mesmo indivíduo (Evans and David Smith, 2015).

RM em dados sumarizados

Essa abordagem é o foco desse trabalho e é um caso particular de aplicação da RM em duas amostras. Muitas vezes pode ser difícil obter dados observados no nível dos indivíduos e com informações suficientes sobre a associação de variáveis genéticas e a exposição e/ou desfecho de interesse. Além disso, muitos instrumentos genéticos, mesmo que estejam associados com a exposição de interesse, explicam apenas uma pequena proporção da sua variação sendo necessário um grande tamanho amostral para conferir poder suficiente para análise por RM (Burgess et al., 2015b).

A RM com dados sumarizados permite combinar resultados já publicados em estudos anteriores, tais como associações descritas em um estudo GWAS de uma população específica, de GWAS combinado de diversas populações/coortes, ou a partir de uma metanálise de GWAS. A Tabela 4 lista alguns exemplos de consórcios GWAS que disponibilizaram dados, geralmente na forma de coeficientes (β) referentes às associações encontradas entre SNPs e exposições e/ou desfechos.

Tabela 4. Exemplos de dados publicados a partir de consórcios de GWAS.

Exposição/Desfecho	**Consórcio**	**Site**
Depressão	Psychiatric Genomics Consortium (PGC)	http://www.med.unc.edu/pge/downloads
Diabetes tipo 2	Diabetes Genetics Replication and Meta-analysis (DIAGRAM)	http://diagram-consortium.org/downloads.html
Doença arterial coronariana	Coronary Artery Disease Genome wide Replication and Meta-Analysis (CARDIOGRAM)	http://www.cardiogrampluscd4.org
Educação	Social Science Genetics Association Consortium (SSGAC)	http://ssgac.org/Data.php
Índice de massa corporal	Genetic Investigation of Anthropometric Traits (GIANT)	http://www.broadinstitute.org/collaboration/giant
Índices Glicemia/Insulina	Meta-Analysis of Glucose and Insulin-related traits (MAGIC)	https://www.magicinvestigators.org/downloads/

É importante ressaltar que as associações G-X e G-Y utilizadas devem ser estimadas por meio de amostras de indivíduos diferentes, porém realizados em populações semelhantes principalmente quanto à idade, etnia e proporção dos sexos. Além disso, os coeficientes

devem ser estimados em relação à adição do mesmo alelo variante do SNP (Haycock et al., 2016).

As duas abordagens mais utilizadas para estimação do efeito causal com o uso de dados agregados são o método ponderado pelo inverso da variância e o método baseado na máxima verossimilhança, os quais serão descritos com mais detalhes na próxima seção.

3.2. Estimação do efeito causal utilizando dados individuais

Nesta seção serão descritos de forma geral e resumida os principais métodos, com aplicações mais específicas no contexto da aplicação da RM em dados sumarizados, foco desse trabalho. Detalhes sobre a estimação do efeito causal e o uso desses métodos no contexto de dados com informações no nível de indivíduo podem ser obtidas em Burgess and Thompson (2015).

As medidas de associação utilizadas nos cálculos para estimação do efeito causal se baseiam, com maior frequência, nas estimativas dos coeficientes provenientes do uso de regressão linear para desfechos contínuos ou de regressão logística para desfechos binários. Existem outras abordagens, tais como modelos paramétricos não lineares ou modelos não paramétricos (Burgess et al., 2015b), as quais não serão apresentadas neste trabalho.

3.2.1 Método de Wald

Também chamado de método da razão dos coeficientes, é o método mais simples para estimação do efeito causal de uma exposição sobre um desfecho. Consiste na razão entre o coeficiente estimado da regressão de Y em relação à G e o coeficiente estimado da regressão de X em relação à G (1), escrita por

$$\hat{\beta}_{IV} = \frac{\widehat{\beta}_{Y|G}}{\widehat{\beta}_{X|G}}, \tag{1}$$

em que o coeficiente $\hat{\beta}_{IV}$ representa o efeito causal de X em Y, $\hat{\beta}_{Y|G}$ é o coeficiente de regressão de Y em relação à G, e $\hat{\beta}_{X|G}$ é o coeficiente da regressão de X em relação à G. Essa abordagem permite usar apenas uma variante genética como variável instrumental. Burgess e Thompson (2015) descrevem o método com detalhes, nos contextos em que a resposta Y pode ser continua ou dicotômica, utilizando uma variável instrumental G dicotômica, politômica ou continua. Por estar limitada ao uso de apenas uma variante genética, esta abordagem não é prioridade e não será utilizada neste trabalho.

3.2.2 Método 2SLS (*Two Stage Least Squares*)

O método de regressão por mínimos quadrados de dois estágios permite a estimação do efeito causal considerando o uso simultâneo de variantes genéticas múltiplas como variáveis instrumentais (VI). O primeiro estágio consiste na regressão do da variável X considerando as VI como exposição, e o segundo estágio, na regressão de Y nos valores preditos de X, resultantes do primeiro estágio. A estimativa de efeito causal é o coeficiente de regressão do segundo estágio e pode ser interpretado como a mudança no desfecho causada pela mudança de uma unidade na exposição X (Burgess and Thompson, 2015; Haycook et al., 2016).

Quando mais de uma variável instrumental é considerada (isto é, duas ou mais variantes genéticas) a estimativa do efeito causal é definida como uma média ponderada das estimativas de associação causal de cada variável instrumental, estimadas separadamente, em que os pesos são determinados pelas estimativas de associação das VI no primeiro estágio.

Como descrito por Burgess e Thompson, considere que existem K variantes genéticas $G_1, G_2, \cdots, G_K$, um desfecho quantitativo Y e uma exposição X. Assim, para os indivíduos indexados por $i = 1,2, \cdots, N$, o modelo de regressão do primeiro estágio especifica que

$$x_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} + \varepsilon_{X_i}.$$

Os valores estimados $\hat{x}_i = \hat{\alpha}_0 + \sum_{k=1}^{K} \hat{\alpha}_k g_{ik}$ são utilizados no modelo de regressão da segunda etapa, definido por

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \varepsilon_{Y_i}.$$

Os erros aleatórios ε_{X_i} e ε_{Y_i} são independentes e têm distribuição normal se ambos os modelos são estimados por mínimos quadrados em duas etapas sucessivas. O efeito causal que se deseja estimar é o coeficiente de regressão β_1. O estimador pontual $\hat{\beta}_1$ é não viesado para β_1, mas a estimativa do erro padrão de $\hat{\beta}_1$ obtida no segundo estágio não está correta, pois ignora a incerteza do primeiro estágio. Assim, pressupondo homocedasticidade do erro aleatório, o modelo pode ser reescrito como

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \varepsilon'_{Y_i},$$

e a variância assintótica do estimador $\hat{\beta}_1$ pelo método 2SLS é igual a $\hat{\sigma}^2 \left(\hat{X}' \hat{X}\right)^{-1}$, em que $\hat{\sigma}^2$ é a variância do erro aleatório ε'_{Y_i}. Entretanto, o método é sensível à violação da

homocedasticidade dos resíduos e, também, das suposições do modelo especificado (como não linearidade, por exemplo). Também é importante destacar que a média e a variância do estimador pelo método 2SLS somente estão definidas quando são usadas ao menos 3 variáveis instrumentais (Burgess and Thompson, 2015).

Para um desfecho Y dicotômico, pode ser usado um modelo de regressão log-linear ou logístico no segundo estágio. O ajuste deste modelo de forma sequencial (isto é, os modelos são estimados em duas etapas sequenciais, de forma separada), tende a subestimar os erros padrão do modelo do segundo estágio, visto que não incorporam a variabilidade do modelo de regressão estimado no primeiro estágio. Uma segunda limitação desta abordagem é a possibilidade de os resíduos do modelo do segundo estágio estarem correlacionados com as VIs, comprometendo a interpretação e validade das estimativas. Outras limitações do método podem ser vistas em Burgess e Thompson (2015). Estes autores também apresentam e discutem o método de dois estágios ajustado (*2SRI-Two Stage Residual Inclusion*), que inclui no modelo do segundo estágio os resíduos do modelo estimado no primeiro estágio, contudo, esta abordagem também não tem sido recomendada. Os métodos baseados na máxima verossimilhança, discutidos a seguir, são alternativas para superar as limitações mencionadas.

3.2.3 Métodos baseados na verossimilhança

Apesar de úteis em algumas situações, os métodos descritos até aqui não fornecem estimativas de máxima verossimilhança, cujos estimadores são não viesados, eficientes e possuem a propriedade de normalidade assintótica. Os métodos de verossimilhança máxima com informação completa (*FIML-full information maximum likelihood*), de verossimilhança máxima com informação limitada (*LIML-limited information maximum likelihood*) e métodos bayesianos podem ser utilizados para estimação do efeito causal.

Utilizando o mesmo contexto da seção anterior, o modelo de equações simultâneas para a verossimilhança com informação completa é escrito como

$$x_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} + \varepsilon_{X_i}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_{Y_i}$$

em que o vetor de erros aleatórios $\boldsymbol{\varepsilon} = (\varepsilon_X, \varepsilon_Y)'$ tem distribuição normal bivariada com vetor de médias $(0,0)'$ e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$, isto é, $\varepsilon \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. A correlação entre

ε_X e ε_Y é devida ao confundimento, e os parâmetros das duas equações são estimados de forma simultânea. Os intervalos de confiança são estimados utilizando-se a propriedade de normalidade assintótica dos estimadores de máxima verossimilhança (Burgess and Thompson, 2015).

No método de máxima verossimilhança com informação completa, todos os parâmetros são estimados simultaneamente, porém, na prática, o principal parâmetro de interesse é o coeficiente de regressão β_1, que representa o efeito causal. No método de verossimilhança com informação limitada, a função de verossimilhança é maximizada perfilando e substituindo cada um dos outros parâmetros, exceto β_1. Este método é considerado uma contrapartida do método 2SLS, produzindo uma estimativa equivalente do efeito causal. No entanto, as estimativas são sensíveis às violações na homocedasticidade dos resíduos ou das suposições do modelo e, portanto, o método não tem sido recomendado, especialmente se as variantes genéticas são consideradas VIs fracas (Burgess and Thompson, 2015).

Uma abordagem bayesiana pode ser utilizada para estimação dos parâmetros do modelo de equações simultâneas, para ambos os métodos FIMl e LIML. Para os indivíduos $i = 1,2,\cdots,N$, o vetor aleatório $(X_i, Y_i)'$ representa a exposição e o desfecho, sendo modelado por meio da distribuição normal bivariada com vetor de médias $(\xi_i, \eta_i)'$ e matriz de variâncias e covariâncias **Σ.** O modelo especifica que a distribuição da média da exposição, ξ_i, é uma função linear das VI (variantes genéticas) g_{ik}, com $k = 1,2,\cdots,K$, e que a distribuição de η_i é uma função linear da média da exposição. Assim, o modelo por ser escrito como

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \boldsymbol{\Sigma}\right)$$

$$\xi_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik}$$

$$\eta_i = \beta_0 + \beta_1 \xi_i.$$

O parâmetro β_1 representa o efeito causal entre as médias verdadeiras ξ_i e η_i, em vez dos valores medidos no desfecho e na exposição. O modelo pode ser estimado por meio de métodos de Monte Carlo por Cadeias de Markov (*MCMC – Markov Chain Monte Carlo*) e a média ou mediana da distribuição à posteriori sãousadas como estimativas pontuais. Os percentis 2,5% e 97,5% delimitam o intervalo com 95% de confiança. Desse modo, o método tem a vantagem de não depender de suposições sobre a distribuição do estimador, e assim as inferências tendem a ser mais robustas, mesmo utilizando variantes genéticas consideradas VIs fracas (Burgess and Thompson, 2015).

Os métodos de máxima verossimilhança e bayesianos também podem ser usados para estimação de efeito causal quando o desfecho é dicotômico, pressupondo-se uma relação funcional linear, no logito, entre a probabilidade do evento e a exposição. Para o i-ésimo indivíduo, $Y_i = 1$ e $Y_i = 0$ representam, respectivamente, a ocorrência e não ocorrência do evento de interesse, com probabilidades $\pi_i = P(Y_i = 1)$ e $1 - \pi_i = P(Y_i = 0)$. Assim, para o método FIML o modelo de equações simultâneas é escrito como

$$x_i = N(\xi_i, \sigma_X^2)$$

$$y_i \sim Bernoulli(\pi_i)$$

$$\xi_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik}$$

$$ln\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i.$$

As estimativas dos parâmetros são obtidas maximizando a função de verossimilhança conjunta L, dada por

$$L = \prod_{i=1}^{N} \left(\pi_i^{y_i} \left(1 - \pi_i^{1-y_i}\right) \frac{1}{\sqrt{2\pi}\sigma_X^2} \left[exp\left\{ -\frac{1}{\sigma_X^2}(x_i - \xi_i)^2 \right\} \right] \right).$$

Pela abordagem bayesiana, as estimativas dos parâmetros podem ser obtidas pela distribuição a posteriori, via MCMC (Burgess and Thompson, 2015).

Ao contrário do método 2SLS, em que os resíduos estimados pelo modelo do primeiro estágio são usados como preditor linear do modelo no segundo estágio, os métodos de máxima verossimilhança estimam todos os parâmetros simultaneamente. Desta forma, estes métodos tendem a quantificar a incerteza da estimação do efeito causal β_1 de forma mais adequada.

Abordagens semi-paramétricas foram propostas para tornar o modelo mais robusto a violações das suposições do modelo especificado, tais como linearidade. Nestes métodos de estimação, é especificada uma forma paramétrica para modelar o desfecho e a exposição, porém não é necessário especificar a distribuição de probabilidades dos erros aleatórios. Neste contexto, as principais abordagens de estimação incluem o método generalizado dos momentos (*GMM-Generalized Method of Moments*) e os modelos de média estrutural (*SMM-Structural Mean Models*) que, não sendo o foco deste trabalho, não serão tratados (Clarke and Windmeijer, 2010).

3.2.4 Validade dos instrumentos e potenciais limitações

Algumas abordagens estatísticas podem melhorar a eficiência das estimativas causais. Por exemplo, a incorporação de covariáveis que expliquem a variação da exposição, mas que não se correlacionem com as VI nem com o caminho causal entre exposição e desfecho, aumenta a precisão da estimação causal. Em modelos com dois estágios, a covariável deve ser ajustada em ambas as regressões. Essa estratégia geralmente aumenta a eficiência e a precisão da estimativa do efeito causal, porém pode viesar esta estimativa caso a covariável seja escolhida de forma inapropriada ou caso ela esteja no caminho causal entre a exposição e o desfecho (Burgess and Thompson, 2015) .

Em algumas situações, variantes genéticas se qualificam como variáveis instrumentais, mas não explicam uma grande proporção da variação da exposição de interesse. A estatística F (*Cragg-Donald F statistic*) do modelo de regressão da exposição sobre a variável instrumental pode ser utilizada para estimar a força da variante genética como variável instrumental. Usualmente a variante é considerada um instrumento fraco quando o valor da estatística F é menor do que 10. Isso significa que a associação entre a variante genética e a exposição é 'fraca', podendo gerar uma estimativa viesada (devido ao confundimento) do efeito causal, geralmente na direção da associação entre a exposição e o desfecho (Burgess et al., 2013). No entanto, existem controvérsias quanto à classificação de VIs como instrumentos fracos utilizando o critério arbitrário $F < 10$. Uma das principais razões é que, ao contrário do coeficiente de determinação R^2, a estatística F não pode ser interpretada como uma medida da força da VI, pois depende do tamanho amostral. Ainda, o ponto de corte igual a 10 foi determinado no contexto do método 2SLS, podendo não ser relevante ou confiável para outros métodos, como na abordagem semi-paramétrica. Assim, a escolha de VIs a priori, a partir de embasamento biológico, é uma das maneiras mais eficientes de se evitar esse tipo de viés (Burgess et al., 2013).

Quando mais de uma variável instrumental é utilizada, pode-se fazer um teste de sobre-indentificação para avaliar se os instrumentos possuem efeitos adicionais sobre o desfecho, além do efeito mediado pela exposição. Em outras palavras, avalia-se se existem associações residuais com o desfecho, que podem ser indicativos de violações das suposições para que as variantes genéticas possam ser consideradas variáveis instrumentais ou, também, se existe uma relação não linear entre a exposição e o desfecho. No entanto, estes testes em geral possuem poder baixo e, assim, têm pouca relevância prática. Este e outros aspectos, como os testes de endogeneidade, são discutidos por Burgess e Thompson (2015), e estão além dos objetivos do trabalho.

3.3 Estimação do efeito causal utilizando dados sumarizados

Diversos consórcios de estudos GWAS (Tabela 4) publicaram resultados de associações de catálogos de variantes genéticas com fatores de risco ou com o status de doença. Estas informações sumarizadas (ou agregadas) podem ser utilizadas para estimar efeitos causais, utilizando a abordagem da RM considerando duas amostras. As estimativas de associação entre as variantes genéticas e a exposição, e entre a exposição e o desfecho, devem ser provenientes de dois conjuntos de dados distintos, ou seja, sem intersecção. Entretanto, para que a estimativa de efeito causal seja válida, é vital que os dois conjuntos de dados representem amostras de uma mesma população (Burgess et al., 2015b).

Para entender o método, considere o contexto em que K variantes genéticas (SNPs), observadas em um único estudo, são utilizadas para estimar o efeito causal de uma exposição com o desfecho. Para todo $k = 1,2,\cdots,K$, X_k agora representa a associação genética entre o k-ésimo SNP e a exposição, cuja estimativa de erro padrão é σ_{X_k}. Similarmente, Y_k agora representa a associação genética entre k-ésimo SNP e o desfecho, com erro padrão estimado σ_{Y_k}.

Para uma exposição quantitativa, a associação genética $X_k$representa a mudança média na exposição para cada aumento de um alelo variante do SNP (0, 1 ou 2). Assim, neste contexto, é representada pelo coeficiente β do modelo de regressão linear da exposição sobre o SNP. Do mesmo modo, para um desfecho quantitativo, Y_ké representada pelo coeficiente β do modelo de regressão linear do desfecho sobre o SNP.

Se a exposição ou o desfecho são dicotômicos, a associação genética X_k ou Y_k, respectivamente, é estimada por meio de regressão logística, por exemplo. Neste caso, estimativas de razões de chances geralmente são publicadas pelos estudos GWAS e, assim, X_k ou Y_k são obtidas por meio da transformação logaritmo natural. Na sequência serão apresentados os dois métodos mais utilizados para estimação de efeitos causais com dados sumarizados.

3.3.1 Método ponderado pelo inverso da variância

Para cada variante genética k $(k = 1,2,\cdots,K)$, estima-se o efeito causal da exposição X sobre o desfecho Y utilizando a razão $\frac{Y_k}{X_k}$. Para a obtenção da estimativa global, estas estimativas são ponderadas pela recíproca da correspondente estimativa da variância $\frac{\sigma_{Y_k}^2}{X_k^2}$, que é estimada de forma aproximada por meio do método delta. A ponderação pelo inverso da

variância é semelhante aos cálculos realizados em uma meta-análise de efeito fixo (Burgess et al., 2013; Burgess et al., 2015b). Assim, a estimativa causal global é representada pela equação (2), abaixo:

$$\widehat{\beta} = \frac{\sum_{k=1}^{K} X_k \frac{Y_k}{\sigma^2_{Y_k}}}{\sum_{k=1}^{K} \frac{X_k^2}{\sigma^2_{Y_k}}}. \quad (2)$$

Pressupondo-se uma relação linear entre exposição e o desfecho, a estimativa $\widehat{\beta}$ representa o aumento médio (ou log-odds) no desfecho causado pelo aumento de uma unidade da exposição, A equação (3) mostra a estimativa aproximada do erro padrão de $\hat{\beta}$, como segue:

$$\hat{\sigma}_{\widehat{\beta}} = \sqrt{\frac{1}{\sum_{k=1}^{K} \frac{X_k^2}{\sigma^2_{Y_k}}}}. \quad (3)$$

O método assume que a razão na equação (2) tem distribuição aproximadamente normal, mas a cobertura do intervalo com 95% de confiança baseado nesta distribuição, calculado utilizando a expressão $\hat{\beta} \pm 1{,}96\,\hat{\sigma}_{\hat{\beta}}$, tende a ser próxima do valor nominal. Se as associações genéticas com a exposição são estimadas com boa precisão, então esta abordagem é considerada uma alternativa simples ao método de máxima verossimilhança, do contrário o último método deve ser utilizado (Burgess and Thompson, 2015).

3.3.2 Métodos baseados na máxima verossimilhança

O modelo pode ser construído pressupondo uma relação linear entre a exposição e o desfecho e uma distribuição normal bivariada para as estimativas de associação genética, ou seja, para todo $k = 1,2,\cdots,K$

$$\begin{pmatrix} X_k \\ Y_k \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \xi_k \\ \eta_k \end{pmatrix}, \begin{pmatrix} \sigma^2_{X_k} & \theta\sigma_{X_k}\sigma_{Y_k} \\ \theta\sigma_{X_k}\sigma_{Y_k} & \sigma^2_{Y_k} \end{pmatrix} \right)$$

Uma relação linear é assumida entre as médias subjacentes de ξ_k e η_k, ou seja, $\eta_k = \beta\,\xi_k$. O coeficiente β representa o efeito causal da exposição X sobre o desfecho Y, assumindo-se que é constante para todas as variantes genéticas $k = 1,2,\cdots,K$. Se Y_k for a

mudança no logito da probabilidade do evento para cada aumento de um alelo, então β é o logaritmo natural da razão de chances. Similarmente, se o logaritmo da probabilidade do evento for modelado, β é o logaritmo natural do risco relativo.

Se as associações X-G e G-Y são estudadas na mesma amostra (amostras sobrepostas), pode-se observar uma possível correlação o entre a exposição e desfecho, representada pelo coeficiente θ, que será igual a 0 se as associações forem estimadas a partir de amostras independentes (Burgess et al., 2013; Burgess et al., 2015b).

O efeito causal β pode ser estimado pelo método de máxima verossimilhança ou por métodos bayesianos, implementados em diferentes softwares, como Stata, SAS, MLwiN ou WinBUG. Para o ambiente R, Burgess et al (2015b) apresentaram rotinas computacionais que serão descritas no próximo capítulo.

É importante salientar que, para combinar estimativas de efeito causal de diferentes variantes genéticas em uma estimativa global é necessário pressupor que as variantes genéticas são independentes. Por exemplo, se as associações forem estimadas utilizando os mesmos dados, então as informações sobre o efeito causal fornecida por cada variante não são independentes. A suposição de independência também pode ser violada se existirem interações entre as variantes genéticas e os fatores de risco (interações gene-gene) ou, ainda, se as distribuições entre as variantes genéticas forem correlacionadas devido à presença de desequilíbrio de ligação (LD). A utilização de SNPs em LD pode superestimar a precisão do intervalo de confiança do efeito de causalidade entre X e Y calculado pelo método ponderado pelo inverso da variância (Burgess et al., 2013). Resultados de estudos de simulação revelaram que a cobertura do intervalo de confiança é menor do que o valor nominal quando as variantes genéticas são correlacionadas (Burgess et al., 2015b).

Se a estimativa de correlação ρ entre os SNPs estiver disponível, o método baseado na verossimilhança pode ser modificado usando uma distribuição normal multivariada para as associações genéticas dos SNPs com X e Y. As estimativas de correlação são incorporadas na matriz de variâncias e covariâncias, sendo que a correlação entre as associações genéticas de dois SNPs é considerada igual à correlação entre os próprios SNPs. As associações genéticas com a exposição X e o desfecho Y são representadas pelos vetores $\boldsymbol{X}_k$ e $\boldsymbol{Y}_k$, para todo $k = 1,2,\cdots,K$, os quais têm distribuição normal multivariada, tais que

$$\boldsymbol{X} \sim N_K(\boldsymbol{\xi}, \boldsymbol{\Sigma_X})$$

$$\boldsymbol{Y} \sim N_K(\beta_L \boldsymbol{\xi}, \boldsymbol{\Sigma_Y}),$$

em que $\boldsymbol{\xi}' = (\xi_1, \xi_2, \cdots, \xi_K)$ e os componentes das matrizes de variâncias e covariâncias são dados por $\Sigma_{X_{ij}} = \rho_{ij}\sigma_{X_i}\sigma_{X_j}$, $\Sigma_{Y_{ij}} = \rho_{ij}\sigma_{Y_i}\sigma_{Y_j}$ e $\Sigma_{X_iY_j} = \theta\rho_{ij}\sigma_{X_i}\sigma_{Y_j}$. Os termos σ_{X_i}, ρ_{ij} e θ representam, respectivamente, o erro padrão associado ao coeficiente X_i, a correlação entre as variantes i e j e correlação entre a associação entre a exposição e o desfecho. O modelo é escrito como

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim N_{2K}\left(\begin{pmatrix} \boldsymbol{\xi} \\ \beta_L\boldsymbol{\xi} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_X} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_Y} \end{pmatrix}\right),$$

e $\boldsymbol{\Sigma_{XY}} = (\boldsymbol{\Sigma_{YX}})'$.

Como no modelo anterior, o parâmetro θ representa a possível correlação X-Y se as associações genéticas foram estimadas utilizando amostras sobrepostas. Caso as estimativas de associação X-G e Y-G tenham sido obtidas por meio de amostras de indivíduos diferentes, então $\theta = 0$. Se estiverem disponíveis dados no nível dos indivíduos, o parâmetro θ pode ser estimado por meio de *bootstraping*, caso contrário é recomendado fazer análise de sensibilidade considerando o conjunto de valores plausíveis. Sintaxes em R para estimação deste modelo foram apresentadas por Burgess et al.,(2015b).

Um cuidado necessário quando se utilizam dados sumarizados é que, em geral, os estudos GWAS indicam um SNP como sendo o principal, isso é, o que apresenta associação mais significativa com a exposição em uma determinada região genética. Utilizar apenas o SNP principal pode acarretar em um viés de seleção, denominado "*Winer's Curse*", uma vez que, nesses casos, a associação G-X em geral é superestimada. Burgess et al., (2013) recomendam escolher os SNPs principais de cada região genética, com base em fontes de dados independentes.

Com dados sumarizados, algumas das principais premissas necessárias para validação da variante genética como variável instrumental geralmente não são satisfeitas. Por exemplo, é muito raro encontrar estudos GWAS que reportem a associação dos SNPs não apenas com o fator de exposição de interesse, mas também com possíveis confundidores. Adicionalmente, a suposição de linearidade das associações genéticas e da associação X-Y também não pode ser verificada quando se utiliza dados já publicados. Sendo assim, é extremamente importante um bom embasamento biológico nas hipóteses pesquisadas por meio da abordagem da RM (Burgess et al., 2013).

Burgess e Thompson (2015) também apresentaram modelos para o contexto de múltiplos estudos, considerando informações no nível do indivíduo ou utilizando dados

sumarizados, bem como uma ou muitas variantes genéticas. Contudo, estas abordagens fogem do escopo deste trabalho.

O próximo capítulo apresenta de forma detalhada um exemplo de aplicação da RM utilizando dados sumarizados, como foi descrito por Taylor et al. (2016). Aspectos computacionais para estimação dos efeitos causais também são detalhados (Burgess et al., 2015b).

4. Aplicação

Este capítulo apresenta uma aplicação de randomização mendeliana. É uma replicação de um exemplo descrito na literatura, que considera dados sumarizados obtidos em duas amostras (estudos distintos) e quatro SNPs não correlacionados. O objetivo deste estudo foi estimar a associação causal entre a concentração sérica de vitamina D e ocorrência de esquizofrenia.

4.1. Deficiência de vitamina D e esquizofrenia (replicação do exemplo)

Esse exemplo foi retirado de Taylor et al. (2016) para demonstração de uma aplicação de RM no contexto de duas amostras com dados sumarizados. O objetivo do estudo é investigar o efeito causal da deficiência na concentração sérica de vitamina D (25(OH)D) em casos de esquizofrenia. Foram utilizados SNPs descritos pelo estudo GWAS SUNLIGHT associados de forma significativa com concentrações séricas de vitamina D em indivíduos de ancestralidade europeia (Tabela 5). Nenhuma variante genética estava em LD.

O estudo SUNLIGHT, no entanto, não fornece informação sobre os coeficientes β e respectivos erros padrão para a associação genética X-G. Assim, essa informação foi obtida de uma metanálise realizada por Vimaleswaran et al., 2013 (Tabela 5).

Tabela 5. Variantes genéticas associadas com a concentração sérica de vitamina D.

SNP	Cr[1]	Gene	*Alelo (Menor/Maior)*	*Frequência do alelo menor*	Efeito/Alelo menor[2]	EP[3]
rs2282679	4	GC	G/T	0,26	-8,45	0,31
rs10741657	11	CYP2R1	A/G	0,40	3,12	0,29
rs12785878	11	DHCR7	G/T	0,22	-3,70	0,30
rs6013897	20	CYP24A1	A/T	0,20	-1,85	0,33

Fonte: Vimaleswaran, KS et al (2013) - Taylor et al., (2016) material suplementar
[1] – Cromossomo
[2] – Mudança em % da concentração sérica de vitamina D para cada aumento de um alelo menor
[3] – Erro padrão

Com respeito à associação genética G-Y, foram utilizados os coeficientes β e os erros padrão provenientes das estimativas de razão de chances fornecidas pelo GWAS conduzido pelo consórcio de psiquiatria genética (*PGC–Psychiatric Genomics Consortium*) (Tabela 6).

Tabela 6. Associação genética dos SNPs relacionados com vitamina D com esquizofrenia.

SNP	Cr[1]	Gene	*Alelo Menor*	RC de Esquizofrenia/Alelo Menor[2]	EP[3]
rs2282679	4	GC	T	0,99730	0,0117
rs10741657	11	CYP2R1	A	0,97775	0,0108
rs12785878	11	DHCR7	T	1,01430	0,0118
rs6013897	20	CYP24A1	A	1,02163	0,0130

Fonte: *Psychiatric Genomics Consortium* - Taylor et al., (2016) material suplementar
[1] – Cromossomo
[2] – Razão de chances de esquizofrenia para cada aumento de um alelo menor
[3] – Erro padrão

A seguir serão descritos os passos para realizar a análise de RM pelo método de ponderação pelo inverso da variância e pelos métodos baseados na verossimilhança.

4.1.1. Método ponderado pelo inverso da variância

1º Passo – Cálculo do efeito $\frac{Y_k}{X_k}$ *para toda variante k*

Inicialmente é necessário verificar se o alelo de efeito (alelo menor) é o mesmo nas duas associações. Para os SNPs rs2282679 e rs12785878 os resultados apresentados consideram alelos contrários como o alelo de efeito. Ainda, é importante notar que a estimativa do efeito G-Y foi reportado pela medida de razão de chances, devendo ser utilizada a transformação logaritmo natural. Assim, para cada SNP, com base nas informações das Tabelas 5 e 6, as estimativas de associações genéticas de Y_k e X_k, e dos respectivos erros padrão σ_{Y_k}e σ_{X_k}, são calculadas da seguinte forma:

SNP: rs2282679

X_1 = 0,0845

Y_1 = log(0,9973)

σ_{X_1} = 0,031

σ_{Y_1} = 0,0117/0.9973 *

*Calculado de forma aproximada, utilizando o método delta: $EP(\beta) = Ep(RC)/\exp(\beta)$.

$$\sigma_{\frac{Y_1}{X_1}} = \sqrt{\frac{1}{\frac{X_1^2}{\sigma_{Y_1}^2}}}$$

SNP: rs10741657

X_2 = 0,0312

Y_2 =log(0,97775)

σ_{X_2} = 0,029

σ_{Y_2} = 0,0108/0,97775

$$\sigma_{\frac{Y_2}{X_2}} = \sqrt{\frac{1}{\frac{X_2^2}{\sigma_{Y_2}^2}}}$$

SNP: rs12785878

X_3 = 0,0370

Y_3 = log(1,0143)

σ_{X_3} =0,030

σ_{Y_3} = 0,0118/1,0143

$$\sigma_{\frac{Y_3}{X_3}} = \sqrt{\frac{1}{\frac{X_3^2}{\sigma_{Y_3}^2}}}$$

SNP: rs6013897

$X_4 = -0{,}0185$

$Y_4 = \log(1{,}02163)$

$\sigma_{X_4} = 0{,}033$

$\sigma_{Y_4} = 0{,}0130/1{,}021632$

$$\sigma_{\frac{Y_4}{X_4}} = \sqrt{\frac{1}{\frac{X_4^2}{\sigma_{Y_4}^2}}}$$

<u>2º Passo</u> – Estimação do efeito causal

As estimativas dos efeitos causais associados a cada SNP e a estimativa do efeito global podem ser obtidas por meio da sintaxe R descrita abaixo:

library(meta)

#vetor do efeito causal estimado para cada SNP, multiplicado por 0.1, pois o interesse é verificar o efeito de 10% de aumento na concentração sérica de vitamina D

effects = c$\left(\frac{Y_1}{X_1} * 0.1, \frac{Y_2}{X_2} * 0.1, \frac{Y_3}{X_3} * 0.1, \frac{Y_4}{X_4} * 0.1\right)$

#vetor do erro padrão calculado para o efeito de cada SNP
see = c($\sigma_{\frac{Y_1}{X_1}} * 0.1, \sigma_{\frac{Y_2}{X_2}} * 0.1, \sigma_{\frac{Y_3}{X_3}} * 0.1, \sigma_{\frac{Y_4}{X_4}} * 0.1$*)*

#Pesos para ponderação (inverso da variância de Y_k
w=c(1/σ_{Y_1}*, 1/*σ_{Y_2}*, 1/*σ_{Y_3}*, 1/*σ_{Y_4}*)*

#Comando para meta-análise de efeito fixo, resultando em uma medida de associação expressa em Odds Ratio ("OR")

rmeta <- metagen(effects, see, sm="OR", comb.random=FALSE)
rmeta

Os resultados são reportados no quadro abaixo:

```
> rmeta
      OR           95%-CI %W(fixed)
1 0.9968 [0.9700; 1.0243]     71.97
2 0.9304 [0.8680; 0.9973]     11.07
3 1.0391 [0.9770; 1.1052]     14.03
4 0.8908 [0.7784; 1.0193]      2.93

Number of studies combined: k=4

                         OR           95%-CI      z  p-value
Fixed effect model 0.9917 [0.9691; 1.0149] -0.704   0.4814

Quantifying heterogeneity:
tau^2 = 0.0016; H = 1.64 [1; 2.82]; I^2 = 62.6% [0%; 87.4%]

Test of heterogeneity:
    Q d.f.  p-value
 8.02    3   0.0455

Details on meta-analytical method:
- Inverse variance method
```

Não há evidência estatística (RC=0,9917; 0,9691–1,0149) de que existe um efeito causal do aumento sérico de vitamina D e a prevalência de esquizofrenia. Os resultados das estimativas dos efeitos individuais e da estimativa combinada dos efeitos podem ser visualizados no ***forest plot*** (Figura 5) obtido por meio da sintaxe R abaixo:

```
forest(rmeta2,studlab=c("rs2282679", "rs10741657", " rs12785878", "rs6013897"), colgap.forest.left=unit(1.5,"inches"), xlab = "OR for schizophrenia per 10% increase in serum vitamin D", weight = 'fixed', ref = 1, leftcols = c("studlab"), leftlabs = c('SNPs'), rightcols= c("effect", "ci"), rightlabs=c("OR", "(95 CI)"), digits=4, digits.pval=3, fs.hetstat=9, fontsize = 12, type.fixed="diamond", print.tau2=FALSE, text.fixed = c("Overall"), addspace=TRUE, hetlab=c(""),squaresize= 1.5, calcwidth.pooled=TRUE, just="center")
```

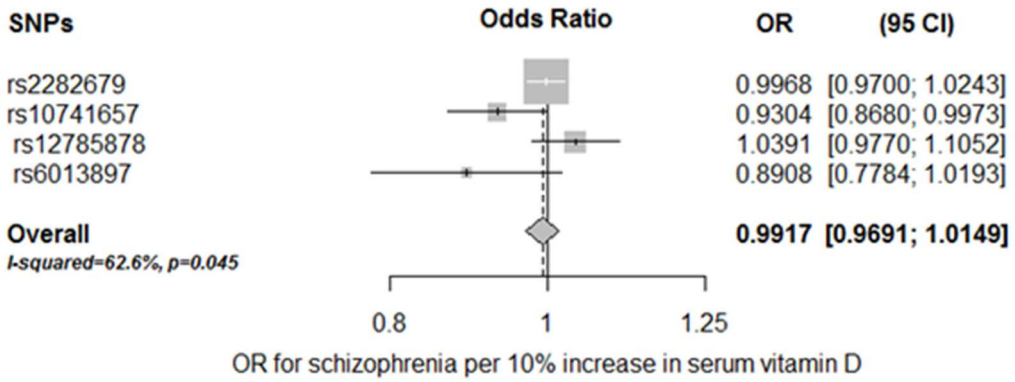

Figura 5 – *Forrest plot* das associações causais dos SNPs com a presença de esquizofrenia estimada pelo método da ponderação pelo inverso da variância.

A estimativa da estatística I^2 (*I-squared*), $I^2 = 62{,}5\%$, indica a presença de alta heterogeneidade entre a associação estimada X-Y, para cada SNP, podendo ser decorrente de pleiotropia, isto é, os SNPs podem não estar associados ao desfecho apenas pelo seu efeito na exposição, mas também podem estar associados a fatores de confundimento, violando uma das premissas para seu uso como variável instrumental.

A presença de pleiotropia pode ser avaliada por meio do modelo de regressão Egger adaptado por Bowden et al., (2015) para o contexto de RM. O método consiste na regressão das estimativas de associações genéticas G-Y em X-G, ponderada pelo inverso da variância de G-Y. Resumidamente, essa regressão é um caso especial do método geral de meta-regressão e, quando aplicada a RM, o intercepto pode ser interpretado como uma estimativa do efeito médio de pleiotropia entre as variantes genéticas. Um intercepto que difere significativamente de zero indica pleiotropia. Além disso, o coeficiente de regressão é considerado uma estimativa válida para estimativa do efeito causal, mesmo quando um ou mais SNPs são pleiotrópicos.

O modelo de regressão de Egger pode ser estimado utilizando os códigos R do quadro abaixo:

x = c(X_1, X_2, X_3, X_4)
y=c(Y_1, Y_2, Y_3, Y_4)
seey = c($\sigma_{Y_1}, \sigma_{Y_2}, \sigma_{Y_3}, \sigma_{Y_4}$)
*BYG = y*sign(x)*
BXG = abs(x)
MREggerFIT = summary (lm(BYG ~ BXG, weights=1/seey^2))
MREggerFIT$coef

No exemplo, os resultados do ajuste do modelo são descritos a seguir, evidenciando que não existe efeito de pleiotropia entre as variantes genéticas utilizadas, haja vista que a hipótese nula H_0: $\beta_0 = 0$ (o intercepto é nulo) não foi rejeitada (p=0,467).

```
> MREggerFIT$coef
               Estimate Std. Error    t value Pr(>|t|)
(Intercept) -0.01790439 0.0201273 -0.8895574 0.4675611
```

4.1.2. Métodos baseados na verossimilhança

Neste exemplo, como as variantes genéticas não estão em desequilíbrio de ligação e são provenientes de amostras distintas, então os parâmetros ρ *e* θ que representam, respectivamente, as correlações entre os SNPs e entre as associações X-G e G-Y podem ser considerados nulos. Neste contexto, os códigos R mostrados no quadro abaixo podem ser usados para maximizar a função de verossimilhança do modelo descrito na Seção 2.3.2.

```
x = c(X1, X2, X3, X4)
y=c(Y1, Y2, Y3, Y4)
seey = c(σY1, σY2, σY3, σY4)
seex = c(σX1, σX2, σX3, σX4)
loglikelihood <- function(param) { # log-likelihood function
        return(1/2*sum((x-param[1:length(x)])^2/seex^2)+1/2*
            sum((y-param[length(x)+1]*param[1:length(x)])^2/seey^2)) }
    opt = optim(c(x, sum(x*y/seey^2)/sum(x^2/seey^2)),
            loglikelihood, hessian=TRUE, control = list(maxit=25000))

# optimization command
cat("Pooled estimate from likelihood-based method: ", opt$par[length(x)+1],
  " \nStandard error: ", sqrt(solve(opt$hessian)[length(x)+1,length(x)+1]))
```

Os resultados desta sintaxe retornam a estimativa global do efeito e erro padrão na escala do logaritmo da razão de chances.

```
Pooled estimate from likelihood-based method:  -0.08353925
Standard error:  0.1182549
```

Assim, a estimativa global de efeito causal é obtida por meio da a exponencial do valor estimado. Similarmente, utilizando a propriedade de normalidade assintótica dos estimadores de máxima verossimilhança, uma estimativa aproximada do IC 95% para o efeito causal pode ser construída usando a exponencial dos limites inferior e superior de confiança na escala do logaritmo da RC. A sintaxe abaixo mostra como isto pode ser obtido por meio do programa R:

```
exp(-0.0835)
exp((-0.0853 - 1.96*0.118))
exp((-0.0853 + 1.96*0.118))
```

A estimativa do efeito causal (RC=0,992; IC 95%:0,969-1,015) por esse método foi semelhante às estimativas anteriores, não sendo encontrada evidência estatística de que a associação entre vitamina D e esquizofrenia seja causal.

A sintaxe R abaixo pode ser utilizada para a maximização da função verossimilhança do modelo utilizando uma abordagem bayesiana.

```
model {
        beta1 ~ dnorm(0, 0.000001)        # priori para o efeito causal
        for (k in 1:K) {                  # para K variants genéticas
        xi[k] ~ dnorm(0, 0.000001)        # priori para a média da associação X-G
        taux[k] <- pow(sigmax[k], -2)     # variação da associação X-G
        tauy[k] <- pow(sigmay[k], -2)     # variação da associação G-Y
        x[k] ~ dnorm(xi[k], taux[k])      # distribuição normal para associação X-G
        eta[k] <- beta1*xi[k]             # média da associação G-Y
        y[k] ~ dnorm(eta[k], tauy[k])     # distribuição normal para associação G-Y
        }
}
# Considerando 4 variantes genéticas (K=4)
x = c(X_1, X_2, X_3, X_4)
y=c(Y_1, Y_2, Y_3, Y_4)
seey = c(σ_Y1, σ_Y2, σ_Y3, σ_Y4)
seex = c(σ_X1, σ_X2, σ_X3, σ_X4)
data <- list ("x", "y", "seey", "seex", "K")
inits <- function() {list (beta1=0)}
parameters <- c("beta1")
summarized.sim <- bugs (data, inits, parameters,
"C:/Users/Paula/Desktop/MR/model.txt", n.chains=1, n.burnin=1000, n.iter=11000,
n.sims=10000, bugs.directory="C:/Users/Paula/Desktop/MR/WinBUGS14")
cat("Pooled estimate from likelihood-based method: ",
summarized.sim$summary[1,1],
   " \n95% credible interval: ", summarized.sim$summary[1,3], ", ",
summarized.sim$summary[1,7])
# Estimativas de efeito (IC 95%)
exp(-0.318)
exp(-0.0848)
exp(0.149)
```

Considerando a escala do logaritmo natural da razão de chances, a estimativa global de efeito causal e o correspondente intervalo de credibilidade com 95% de probabilidade foram:

```
Pooled estimate from likelihood-based method:  -0.08483104
95% credible interval:  -0.3178275 ,  0.1491
```

Exponenciando os valores obtidos, vê-se que a estimativa global de efeito causal RC=0,992, com intervalo de credibilidade com 95% de probabilidade (0,969-1,015), foi semelhante às estimativas obtidas métodos descritos anteriormente. Portanto, não foi encontrada evidência estatística de que a associação causal entre vitamina D e esquizofrenia.

5. Considerações Finais

A Randomização Mendeliana é um método robusto que pode ser utilizado para investigar se a associação observacional entre uma exposição e desfecho de interesse possui natureza causal. O descobrimento de variantes genéticas associadas a exposições modificáveis, intermediários biológicos, e mediadores epigenéticos vem aumentando consideravelmente, aumentando a possibilidade de utilização dessa técnica.

Exposições que antes não poderiam ser avaliadas por ensaios clínicos devido a questões éticas, tais como dietas prejudiciais, álcool, tabaco e inclusive exposição à poluição, podem ser avaliadas em um estudo GWAS de larga escala, em uma população que esteja exposta a esses fatores de risco, revelando variantes genéticas válidas para serem utilizadas como variáveis instrumentais. Com dados disponíveis dos grandes consórcios a RM em dados sumarizados se torna uma aliada na investigação desses fatores de risco com doenças crônicas complexas.

Existem hoje, também, ferramentas online que resumem as associações genéticas observadas para um grande número de SNPs, referenciando o estudo GWAS original: o GWAS Catalog (https://www.ebi.ac.uk/gwas/), GWAS Central (http://www.gwascentral.org/) e SNPedia (http://www.snpedia.com/) são os principais exemplos.

A metodologia de RM, apesar de suas limitações, tem como vantagem ser capaz de auxiliar na descoberta de novos alvos para intervenção terapêutica. A RM possibilita a realização de inferências causais, mas a magnitude estimada de efeito não pode ser avaliada apenas de forma literal, sendo a direção estimada do efeito causal a informação mais relevante. Isso porque as variantes genéticas utilizadas, além de no geral representarem uma pequena parcela da variação da exposição, também podem estar representando apenas um dos possíveis caminhos causais entre exposição e desfecho.

Por ser uma abordagem relativamente nova, essas limitações ainda são tópicos relevantes de estudos e em um futuro recente métodos estatísticos cada vez mais robustos devem ser propostos.

Os conceitos e métodos apresentados neste trabalho constituem apenas a porta de entrada para o estudo da randomização mendeliana e aplicações, um tema complexo e que tem sido extensamente descrito na literatura. Entretanto, por ser escrito em língua portuguesa, pode contribuir divulgação do método, bem como ajudar os leitores iniciantes.

Por fim, é importante salientar as potencialidades da randomização mendeliana em aplicações de diversas áreas, utilizando dados sumarizados já disponíveis na literatura, podendo gerar ou comprovar novas hipótese científicas de forma rápida e baixo custo.

6 Referências

Alberts, B. (2004). Biologia celular interativa (Porto Alegre (RS): Artes Medicas).

Allard, C., Desgagné, V., Patenaude, J., Lacroix, M., Guillemette, L., Battista, M., Doyon, M., Ménard, J., Ardilouze, J., Perron, P., et al. (2015). Mendelian randomization supports causality between maternal hyperglycemia and epigenetic regulation of leptin gene in newborns. Epigenetics *10*, 342–351.

Barrett-Connor, E., Slone, S., Greendale, G., Kritz-Silverstein, D., Espeland, M., Johnson, S.R., Waclawiw, M., and Fineberg, S.E. (1997). The Postmenopausal Estrogen/Progestin Interventions Study: primary outcomes in adherent women. Maturitas *27*, 261–274.

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int. J. Epidemiol. *44*, 512–525.

Burgess, S., and Thompson, S.G. (2015). Mendelian randomization: methods for using genetic variants in causal estimation. CRC Press.

Burgess, S., Butterworth, A., and Thompson, S.G. (2013). Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data: Mendelian Randomization Using Summarized Data. Genet. Epidemiol. *37*, 658–665.

Burgess, S., Small, D.S., and Thompson, S.G. (2015a). A review of instrumental variable estimators for Mendelian randomization. Stat. Methods Med. Res.

Burgess, S., Consortium, E.-I., Scott, R.A., Timpson, N.J., Davey Smith, G., and Thompson, S.G. (2015b). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur. J. Epidemiol. *30*, 543–552.

C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) (2011). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ *342*, d548–d548.

Clarke, P.S., and Windmeijer, F. (2010). Identification of causal effects on binary outcomes using structural mean models. Biostatistics *11*, 756–770.

Clarke, R., Peden, J.F., Hopewell, J.C., Kyriakou, T., Goel, A., Heath, S.C., Parish, S., Barlera, S., Franzosi, M.G., Rust, S., et al. (2009). Genetic Variants Associated with Lp(a) Lipoprotein Level and Coronary Disease. N. Engl. J. Med. *361*, 2518–2528.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. Science, *300*(5617), 286-290.

Davey, D.A. (2012). Update: estrogen and estrogen plus progestin therapy in the care of women at and after the menopause. Womens Health *8*, 169–189.

Davey Smith, G., and Ebrahim, S. (2003). “Mendelian randomization”: can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. *32*, 1–22.

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. *23*, R89–R98.

Didelez, V., and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. Stat. Methods Med. Res. *16*, 309–330.

Evans, D.M., and Davey Smith, G. (2015). Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. Annu. Rev. Genomics Hum. Genet. *16*, 327–350.

Glennan, S. (1996). Mechanisms and the nature of causation. Erkenntnis *44*.

Gray, R., and Wheatley, K. (1991). How to avoid bias when comparing bone marrow transplantation with chemotherapy. Bone Marrow Transplant. *7 Suppl 3*, 9–12.

Greenland, S., and Brumback, B. (2002). An overview of relations among causal modelling methods. Int. J. Epidemiol. *31*, 1030–1037.

Greenland, S., Pearl, J., and Robins, J.M. (1999). Causal diagrams for epidemiologic research. Epidemiol. Camb. Mass *10*, 37–48.

Haycock, P.C., Burgess, S., Wade, K.H., Bowden, J., Relton, C., and Davey Smith, G. (2016). Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am. J. Clin. Nutr. *103*, 965–978.

Hernan, M.A. (2004). A definition of causal effect for epidemiological research. J. Epidemiol. Community Health *58*, 265–271.

Hernán, M.A., and Robins, J.M. (2006). Instruments for causal inference: an epidemiologist's dream? Epidemiol. Camb. Mass *17*, 360–372.

Hill, A.B. (1965). THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? Proc. R. Soc. Med. *58*, 295–300.

Holland, P.W., Glymour, C., and Granger, C. (1985). STATISTICS AND CAUSAL INFERENCE*. ETS Res. Rep. Ser. *1985*, i-72.

Hoppe, A.A., and Carey, G.B. (2007). Polybrominated Diphenyl Ethers as Endocrine Disruptors of Adipocyte Metabolism**. Obesity *15*, 2942–2950.

Hunter, D.J. (2005). Gene-environment interactions in human diseases. Nat. Rev. Genet. *6*, 287–298.

Katan, M.B. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet Lond. Engl. *1*, 507–508.

Katan, M.B. (2004). Commentary: Mendelian randomization, 18 years on. Int. J. Epidemiol. *33*, 10–11.

King, R.C. , Mulligan, P.K., and Stansfield, W.D., (2013) A dictionary of genetics. Oxford University Press.

Lang, T., Klein, K., Fischer, J., Nüssler, A.K., Neuhaus, P., Hofmann, U., Eichelbaum, M., Schwab, M., and Zanger, U.M. (2001). Extensive genetic polymorphism in the human CYP2B6 gene with impact on expression and function in human liver. Pharmacogenetics *11*, 399–415.

Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association

analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.

Palmer, T.M., Nordestgaard, B.G., Benn, M., Tybjaerg-Hansen, A., Davey Smith, G., Lawlor, D.A., and Timpson, N.J. (2013). Association of plasma uric acid with ischaemic heart disease and blood pressure: mendelian randomisation analysis of two large cohorts. BMJ *347*, f4262–f4262.

Parascandola, M. (2001). Causation in epidemiology. J. Epidemiol. Community Health *55*, 905–912.

Penell, J., Lind, L., Fall, T., Syvänen, A.-C., Axelsson, T., Lundmark, P., Morris, A.P., Lindgren, C., Mahajan, A., Salihovic, S., et al. (2014). Genetic variation in the CYP2B6 Gene is related to circulating 2,2',4,4'-tetrabromodiphenyl ether (BDE-47) concentrations: an observational population-based study. Environ. Health *13*.

Robins, J.M., Henán M.A., and Brumback B. (2000). Marginal scructural models and causal inference in epidemiology. Epidemiology 11(5), 550-60.

Rossouw, J.E., Anderson, G.L., Prentice, R.L., LaCroix, A.Z., Kooperberg, C., Stefanick, M.L., Jackson, R.D., Beresford, S.A.A., Howard, B.V., Johnson, K.C., et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA *288*, 321–333.

Rothman, K.J., and Greenland, S. (2005). Causation and Causal Inference in Epidemiology. Am. J. Public Health *95*, S144–S150.

Rothman, K.J., Greenland, S., and Lash, T.L. (2008). Modern epidemiology (Philadelphia Baltimore New York London Buenos Aires Hong Kong Sydney Tokyo: Wolters Kluwer Health, Lippincott Williams & Wilkins).

Sanson-Fisher, R.W., Bonevski, B., Green, L.W., and D'Este, C. (2007). Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions. Am. J. Prev. Med. *33*, 155–161.

Taylor, A.E., Burgess, S., Ware, J.J., Gage, S.H., Richards, J.B., Davey Smith, G., and Munafò, M.R. (2016). Investigating causality in the association between 25(OH)D and schizophrenia. Sci. Rep. *6*, 26496.

Thompson, J.R., Minelli, C., Abrams, K.R., Tobin, M.D., and Riley, R.D. (2005). Meta-analysis of genetic studies using Mendelian randomization—a multivariate approach. Stat. Med. *24*, 2241–2254.

Vimaleswaran, K.S., Berry, D.J., Lu, C., Tikkanen, E., Pilz, S., Hiraki, L.T., Cooper, J.D., Dastani, Z., Li, R., Houston, D.K., et al. (2013). Causal Relationship between Obesity and Vitamin D Status: Bi-Directional Mendelian Randomization Analysis of Multiple Cohorts. PLoS Med. *10*, e1001383.

Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., Johnson, T., et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. The Lancet *380*, 572–580.

Wahl, M., Lahni, B., Guenther, R., Kuch, B., Yang, L., Straehle, U., Strack, S., and Weiss, C. (2008). A technical mixture of 2,2′,4,4′-tetrabromo diphenyl ether (BDE47) and brominated

furans triggers aryl hydrocarbon receptor (AhR) mediated gene expression and toxicity. Chemosphere *73*, 209–215.

Wannmacher, L., and Lubianca, J.N. (2004). Terapia de reposição hormonal na menopausa: evidências atuais. Uso Racion. Medicam. Temas Selecionados *1*, 1–6.

Wium-Andersen, M.K., Ørsted, D.D., and Nordestgaard, B.G. (2014). Elevated C-Reactive Protein, Depression, Somatic Diseases, and All-Cause Mortality: A Mendelian Randomization Study. Biol. Psychiatry *76*, 249–257.

Ye, Z., Sharp, S.J., Burgess, S., Scott, R.A., Imamura, F., Langenberg, C., Wareham, N.J., and Forouhi, N.G. (2015). Association between circulating 25-hydroxyvitamin D and incident type 2 diabetes: a mendelian randomisation study. Lancet Diabetes Endocrinol. *3*, 35–42.

Hans-Olov, A. (2008). Textbook of Cancer Epidemiology. Oxford University Press.

Single nucleotide polymorphism / SNP | Learn Science at Scitable (http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295) Acessado em: 20/11/2016.