Tomada de Decisão Baseada em Evidência

Como julgar o nível de confiança e o grau de recomendação de artigos científicos na área das ciências da saúde

Wellington Lunz

Tomada de Decisão Baseada em Evidência

Como julgar o nível de confiança e o grau de recomendação
de artigos científicos na área das ciências da saúde

Wellington Lunz

1ª edição
Vitória – ES, 2025
ISBN: 978-65-01-57126-3

Projeto gráfico, diagramação, capa e revisão: Wellington Lunz

Publicado por Wellington Lunz. Vitória-ES, 2025

Dados Internacionais de Catalogação na Publicação (CIP)
Câmara Brasileira do Livro, SP, Brasil)

Lunz, Wellington

Tomada de decisão baseada em evidência [livro eletrônico] : como julgar o nível de confiança e o grau de recomendação de artigos científicos na área das ciências da saúde / Wellington Lunz. -- Vitória, ES : Ed. do Autor, 2025.

PDF

Bibliografia.

ISBN 978-65-01-57126-3

25-284490 CDD-610.3

Índices para catálogo sistemático:

1. Ciências da saúde: Pesquisa 610.3

Eliete Marques da Silva - Bibliotecária - CRB-8/9380

Tamanho de efeito (e seu intervalo de confiança 95%)

Uma das coisas mais associadas ao benefício da intervenção e mais fundamentais para a tomada de decisão é o tamanho do efeito. Ou seja, o quão expressivo é o resultado.

Já enfatizei que resultados estatisticamente significativos não denotam necessariamente resultados com magnitudes convincentes para uma tomada de decisão. Temos uma infinidade de exemplos na ciência. Então, como avaliar o valor do efeito?

Na maioria das subáreas da saúde, penso que a estratégia mais usual é considerar o tamanho de efeito (TE). Mas também é fundamental considerar o intervalo de confiança 95% (IC95%) desse TE. Isso porque o TE isoladamente não considera o tamanho amostral, enquanto o IC95% sim.

O que é o IC95% e sua importância, você já aprendeu no capítulo sobre nível de evidência, quando abordei o quesito 'Medidas são precisas?'

O TE permite enxergar a magnitude relativa, por ser padronizada, oferecendo mais condições para uma aplicação prática e clínica.

O TE também está associado positivamente ao poder estatístico (probabilidade de afirmar que algo é verdadeiro quando de fato é; análogo ao chamado 'verdadeiro-positivo') (Lakens, 2013; Espirito Santo e Daniel., 2015).

Quem popularizou o conceito de TE (*effect size*) e mais contribuiu com equações para isso, foi o estatístico e psicólogo Jacob Cohen. O TE mais amplamente conhecido é exatamente o 'd de Cohen'.

Os TE mais usados nas subáreas da saúde pertencem à **família d** e à **família r** (Lakens, 2013; Espirito Santo e Daniel., 2015). Nas áreas epidemiológicas e clínicas, as medidas de TE mais comuns são o risco relativo (RR) e a razão de chances (*odds ratio*, OR).

Vale destacar que essas medidas de efeito podem ser convertidas entre elas, com equações específicas. Pode-se, por exemplo, converter 'r' em 'd' (e vice-versa).

Vou iniciar com a família d, que é a mais amplamente usada:

As equações da família d possuem o mesmo **numerador**, sendo basicamente a diferença (subtração) entre a média do 'grupo intervenção' pelo 'grupo controle' (média 1 – média 2).

Por sua vez, o **denominador** muda entre as diferentes equações da família d. O 'd de Cohen' usa no denominador basicamente a média dos desvios padrões dos dois grupos (intervenção e controle).

Na prática, essa medida nos mostra o quão distante uma média está da outra em termos de desvios padrões. Trata-se, portanto, de uma medida padronizada, porque relativiza as diferenças de médias pela média dos desvios padrões dos dois grupos.

Veja um exemplo didático: suponha dois grupos, sendo que um com média 4 e o outro com média 2. Para facilitar, suponha que ambos os grupos têm desvio padrão igual a 1. Portanto, a média dos desvios padrões será 1. O cálculo seria o seguinte: (4 – 2)/1 = 2.

Repare nas representações das curvas normais da Figura 12 que a distância entre as médias (linhas verdes) dos dois grupos dá exatamente 2. Essa distância representa duas unidades de desvio padrão (DP = 1).

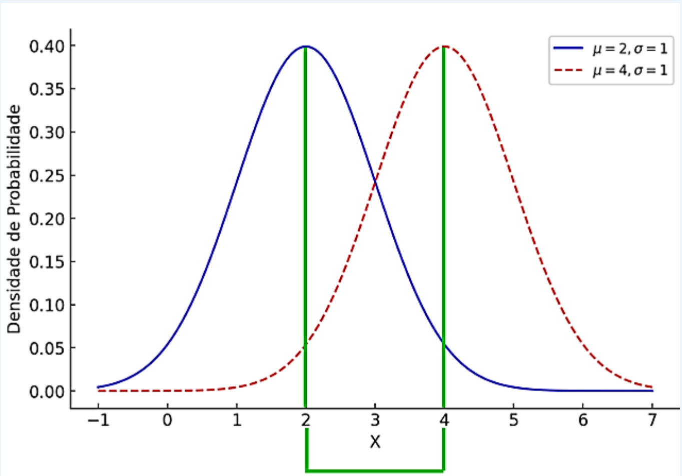

Figura 12. Ilustração do tamanho de efeito a partir da equação de Cohen (d de Cohen).

Quando o tamanho amostral dos grupos é diferente, o denominador será uma média ponderada dos desvios padrões.

Se o artigo científico que você estiver lendo não apresentar o TE, há calculadoras online para isso. Mas recomendo calcular usando as planilhas oferecidas por Lakens (2013) ou Espirito Santo e Daniel (2015) (ou outros artigos científicos que ofereçam planilhas). Você poderá acessar tais artigos e suas planilhas seguindo o passo a passo da ilustração abaixo.

Clique no *QR code* para acessar os links dos artigos. Considero a primeira referência (de Espirito Santo e Daniel., 2015) mais simples e intuitiva (inclusive está em português). Você conseguirá fazer *download* dos artigos e das planilhas (veja a primeira seta indicando os arquivos):

Figura 13. Orientação de como acessar os artigos e planilhas de Espirito Santo e Daniel (2015) e Lakens (2013).

Caso tenha dificuldade, as referências também estão citadas no livro. E se ainda assim não consegui, faça contato comigo (via site: wellingtonlunz.com.br), que repasso os arquivos. Se você ler ambos os artigos, também terá uma compreensão mais ampla sobre TE.

Ainda na família d de TE, há o ‘g de Hedges’, que basicamente faz uma correção do d de Cohen para amostras pequenas. E há o ‘delta de Glass’, que propõe usar o desvio padrão do grupo controle quando as variâncias não são homogêneas.

Mas é importante destacar que o TE também tem limitações. A mais importante é não considerar o tamanho amostral. Por isso, é sempre muito importante verificar o IC95% do TE, o qual considera o tamanho amostral. As planilhas que indiquei também calculam o IC95%.

Se o intervalo de confiança for grande, é porque a amostra é pequena. E isso diminui a precisão e a confiança no TE.

Outro problema é que nem sempre um TE grande representa um resultado que possa ser expressivo para você. E nesse momento vale a pena discutir a classificação do TE.

O mais usual, apesar das críticas, tem sido considerar a sugestão original de Cohen, onde um TE igual a 0,2 é considerado pequeno, 0,5 seria médio e 0,8 seria grande.

Outra sugestão é considerar TE <0,19 como insignificante, de ‘0,2 a 0,49’ como pequeno, de ‘0,50 a 0,79’ como médio, de ‘0,80 a 1,29’ como grande e >1,30 como muito grande (Espirito Santo e Daniel., 2015).

Mas, dependendo do que se está medindo (ex: taxa de suicídio; redução de mortalidade), um TE considerado insignificante ou pequeno por essas classificações poderia ser interpretado como grande.

Em outras vezes, algo >0,8 pode não representar algo importante. No próximo tópico, descreverei a importância de também considerar a magnitude absoluta, exatamente porque o TE isolado pode ser insuficiente.

Outro problema do TE é que se trata de uma medida de distância na escala de 'desvio padrão', de modo que muitos têm dificuldade de entender o que seria um TE relevante. Ou seja, não é uma medida do dia a dia que se possa compreender facilmente.

Para diminuir esse problema, gosto da chamada '**probabilidade de superioridade**', proposta por McGraw e Wong (1992), que também é conhecida como 'TE de linguagem comum'.

Vou explicar agora, mas, se tudo lhe parecer muito difícil, não precisa se preocupar com os cálculos, porque as planilhas dos artigos que já citei (Lakens, 2013; Espirito Santo e Daniel., 2015) também calculam essa 'probabilidade de superioridade'.

A 'probabilidade', em particular quando convertida em percentual, é algo que a maioria das pessoas entende bem. Vamos entender a lógica da proposta:

Suponha que você saia agora em direção à rua e decida medir a estatura do primeiro homem e da primeira mulher que encontrar. Se você tivesse que apostar qual deles será mais alto, você apostaria que seria o homem ou a mulher?

Se você for como a maioria das pessoas, ainda que saiba que existem muitas mulheres mais altas que muitos homens, você apostaria que o homem será mais alto. É algo intuitivo, pois, na média, homens são mais altos. Mas qual é a real probabilidade de você acertar?

McGraw e Wong (1992) usaram exatamente o exemplo das estaturas de homens e mulheres para propor a 'probabilidade de superioridade'. Usaram uma base de dados real, com milhares de homens e mulheres, e as estaturas 'médias (desvio padrão)' foram: homens = 177 cm (7,11 cm) e mulheres = 163 cm (6,60 cm).

Como se percebe, a diferença das médias dá 14 cm. Ou seja, se medirmos par-a-par homens e mulheres aleatórios, espera-se que a média de diferença de estatura seja de 14 cm, e que o desvio padrão seja uma média dos dois desvios. Isso poderia ser descrito por uma curva gaussiana (curva normal), a qual represento na Figura composta abaixo.

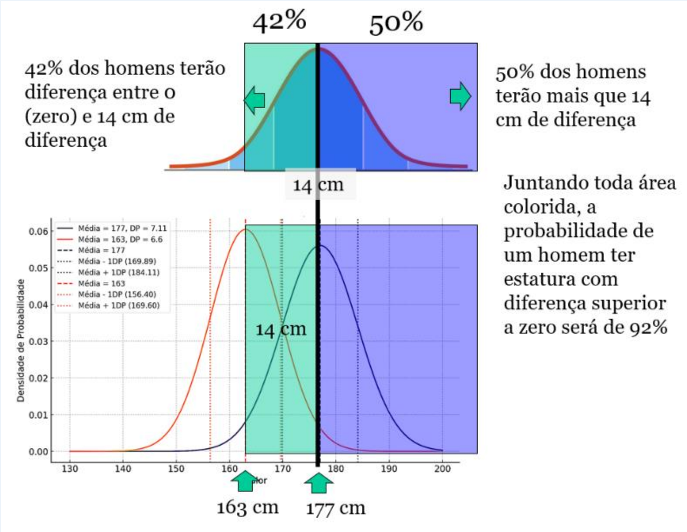

Figura 14. Ilustração de como interpretar a probabilidade de superioridade proposta por McGraw e Wong (1992).

Se considerarmos a curva normal da Figura 14, e considerando que a diferença média de estatura é 14 cm, então teremos que 50% dos homens (metade da curva normal) terão estatura acima de 14 cm de diferença.

Mas qual a probabilidade de encontrar homens com diferença menores que 14 cm até 0 (zero) cm em relação às mulheres?

A equação de McGraw e Wong (1992) calcula exatamente essa probabilidade, a qual se soma aos demais 50% que descrevi no parágrafo anterior.

E o que McGraw e Wong (1992) encontraram nos dados e exemplos das estaturas foi que a probabilidade de um homem ser maior que uma mulher é de 92%, como ilustrado na Figura 14Figura.

A interpretação é a seguinte: 'ao selecionar aleatoriamente um homem e uma mulher, a probabilidade de um homem ter estatura maior que uma mulher é de 92%'.

Esse resultado dá uma boa noção da magnitude. Se você tivesse que apostar agora, teria uma noção mais clara da sua real chance de ganhar.

E toda essa lógica pode ser usada quando você está comparando dois procedimentos, ou técnicas, ou fármacos, ou treinamento, etc.

Aos mais interessados em se aprofundar no assunto, podem ler o estudo original de McGraw e Wong (1992).

Aliás, embora na Figura ilustrativa eu já coloquei os valores em percentual, na verdade a equação apresenta o 'z-score' da região abaixo da média, o que pode ser complicado para quem não tem familiaridade com estatística científica.

Em resumo, se você não for uma pessoa com boa aproximação com a matemática e estatística científica, recomendo que use a planilha de cálculos dos artigos sugeridos. E, claro, todo esse detalhamento é mais direcionado para os aspirantes a cientistas. Para os leitores de artigos, basta a compreensão geral. E darei mais exemplo prático para ajudar nisso.

Vale também registrar que a equação descrita por Lakens (2013) e Espirito Santo e Daniel. (2015) em seus artigos está errada, mas está correta nas planilhas. A descrição da equação correta também está no próprio artigo original de McGraw e Wong (1992).

Vou agora dar um exemplo operacional do uso da **probabilidade de superioridade**. Vou usar dados aproximados do artigo científico de Kassiano et al. (2024).

Eles objetivaram verificar se os exercícios físicos '*leg press* + *stiff* + elevação pélvica' (grupo LSE) induziriam maior hipertrofia no músculo glúteo máximo se comparados a somente fazer '*leg press* + *stiff*' (grupo LS).

Ou seja, o grupo LSE tinha um exercício a mais, que era exatamente o exercício chamado 'elevação pélvica'.

Após algumas semanas de treino para hipertrofia, viram que o aumento médio da espessura muscular glútea foi de 6% e de 9,3% para os grupos LS e LSE, respectivamente. Para os desvios padrões, tive que fazer uma estimativa a partir das Figuras (DP = 4,9% para ambos os grupos).

Apesar da diferença percentual média a favor do grupo LSE, percebe-se, pela Figura 3 do artigo, que várias pessoas do grupo LS tiveram resultados acima da média do LSE. E, como esperado, um número maior de pessoas do grupo LSE teve resultados maiores que a média do LS. E é por isso que a probabilidade de superioridade é útil para auxiliar nessa interpretação.

Calculando conforme McGraw e Wong (1992), a probabilidade de superioridade a favor do grupo LSE deu 68%.

E como interpretar esse resultado? Veja:

Se pegarmos aleatoriamente duas pessoas, sendo que uma treinou LS e outra treinou LSE, exatamente como no protocolo do estudo, a probabilidade da pessoa que treinou com LSE ter conseguido maior ganho de espessura glútea é de 68%. A probabilidade da pessoa do grupo LS conseguir melhor resultado será de apenas 32% (ou seja, não é pior sempre).

Diante disso, a questão que fica é: essa probabilidade vale a pena para você?

E a resposta é pessoal. Você precisa considerar muitas coisas. Por exemplo, o grupo LSE sempre teve que fazer um exercício a mais, por várias semanas. A ‘elevação pélvica’ é um exercício que exige montar a carga de peso (colocar anilhas na barra). E, como o glúteo é um músculo forte, exigirá várias anilhas. Depois, você tem que desmontar e guardar os pesos.

Você gosta ou não de fazer o exercício ‘elevação pélvica’? E qual foi o ganho absoluto? Só o resultado relativo (em %) não é suficiente para convencer, como mostrarei no próximo tópico.

Só para antecipar essa questão do ganho absoluto: a média de ganho absoluto para o grupo LS foi de 2,5 milímetros (mm). E a média de ganho absoluto para o grupo LSE foi de 4 mm. Comparando ambos os grupos, será que esse esforço adicional do grupo LSE, de fazer um exercício a mais, vale a pena, considerando ganhar, em média, 1,5 mm (4 – 2,5) a mais que o outro grupo? E isso nem acontecerá sempre.

Entrei com esses mesmos dados do estudo científico de Kassiano et al. (2024) na planilha de cálculo de Espirito Santo e Daniel. (2015). O registro da saída (*output*) dos resultados você pode ver na Figura 15 (abaixo).

A seta verde (na Figura) aponta o TE calculado por três estratégias da família d. A seta azul indica o IC95%, o qual só será calculado se for incluído o tamanho amostral de cada grupo (n=10). O t de *Student* e o *p-value* também foram calculados. A seta vermelha indica a probabilidade de superioridade.

	Sujeitos	Média	Desvio padrão
Grupo alvo	10	9,3	4,9
Grupo de comparação	10	6	4,9

	***d* de Cohen**	***g* de Hedges**	**Delta de Glass**
***d* de Cohen, *g*, delta**	0,67	0,65	0,67
Interpretação	*Efeito médio*		
IC 95% TDE [Baixo, Alto]	-0,25	1,54	
IC 95% M_{dif} [Baixo, Alto]	-1,30	7,90	
Teste *t*	1,51		
Graus de liberdade *(gl)*	18		
Nível de significância *(p)*	0,149434		
TDE-LC	68,3		

Figura 15. Ilustração da saída (*output*) de resultados usando a planilha de Espirito Santo e Daniel (2015).

Então, considerando esses dados, temos um TE classificado como médio (TE=0,65 a 0,67), com probabilidade de superioridade de 68% a favor do grupo LSE. O TE baseado na 'família d' tem sido o mais usado, e por isso me dediquei um pouco mais nele.

Embora fuja ao escopo do presente livro, outra boa estratégia para verificar a magnitude de efeito (mas não só isso) usando a linguagem de probabilidade é via análise inferencial bayesiana. A vantagem, comparada ao método tradicional ou frequencista (que considera o *p-value* e uma distribuição estimada), é que a análise bayesiana considera dados prévios (*prior*) reais, e não uma distribuição estimada. Isso permite entender se uma intervenção foi melhor que aquilo que o conhecimento prévio apresentava. Portanto, considere também estudar o teorema de Bayes e análise bayesiana no processo de tomada de decisão.

Em relação à 'família r' de TE, essa tem a ver com correlações e associações entre variáveis de interesse.

Exemplo: sabemos que há relação linear positiva entre o envelhecimento e o ganho de peso corporal, ou entre o consumo de sal de cozinha e a pressão sanguínea, ou entre o consumo calórico e o ganho de gordura, e entre muitas outras coisas.

Há muitas variáveis que se associam e, com bastante frequência (mas não obrigatoriamente), a variação de uma variável explica a variação da outra. A família r quantifica exatamente essa magnitude. Quanto mais forte for a correlação ou associação, maior o efeito.

Na família r destacam-se o r (coeficiente de correlação), o r^2 (coeficiente de determinação) e o eta quadrado (η^2; *eta squared*).

O r é uma medida da força da associação, com valores entre -1 a +1. Os sinais negativo e positivo indicam a direção da associação, e não a força. Quanto mais perto de ±1, mais forte é a correlação; quanto mais perto de 0 (zero), menor é a força.

O r^2 e o η^2 permitem interpretar a força de associação na linguagem de probabilidade. Por exemplo, um r^2 e η^2 de 0,15 permite interpretar que 15% da variação da variável dependente (no caso do r^2) e da variação total (no caso do η^2) pode ser explicada pela variável independente (no caso do r^2) ou pelo efeito da intervenção (no caso do η^2).

Suponha um teste de correlação entre idade e pressão arterial. Caso dê r^2 de 0,10, interpreta-se que a variação na idade pode explicar 10% da variação da pressão arterial.

Isso não significa poder afirmar que o aumento da pressão é (ou foi) explicado somente por alterações fisiológicas associadas a idade. Afinal, à medida que envelhecemos, há também mudanças de comportamentos (ex: sedentarismo, alimentação) que poderiam contribuir.

Agora, suponha que você teste 3 tipos de dietas na redução de gordura corporal, e usou a estatística ANOVA (o η^2 já costuma ser apresentado no *output*). Caso o η^2 do efeito (dietas) dê 0,10, é porque 10% da redução de gordura corporal foi explicada pela intervenção (dieta). Para saber qual das 3 dietas foi mais importante, é preciso complementar usando um teste *post-hoc*.

Se a ANOVA não apresentar o η^2, saiba que é algo fácil de calcular. Uma rápida pesquisa na internet ou usando boas inteligências artificiais, você conseguirá. Ainda mais recomendado é acessar trabalhos científicos, como o de Lakens (2013), o qual explica.

Também não há consenso sobre a classificação da força do r, r^2 e η^2. Há muitas classificações diferentes, mas elas ficam em torno do que descrevo na tabela abaixo.

Tabela 1. Classificação habitual dos coeficientes de correlação (r) e determinação (r^2) e do eta quadrado (η^2).

r (+ ou -)	**Classif.**	**r^2**	**Classif.**	**η^2**	**Classif.**
0-0,19	Muito fraca	<0,04	Muito fraca	<0,01	Despre-zível
0,2-0,39	Fraca	0,05-0,15	Fraca	0,01-0,06	Pequeno
0,4 – 0,69	Modera-da	0,16-0,48	Modera-da	0,07-0,13	Médio
0,7-0,89	Forte	0,49-0,79	Forte	≥0,14	Grande
0,9-1	Muito forte	0,80-1	Muito forte		

Legenda: Classif. = classificação

A falta de consenso sobre essa classificação é porque esse julgamento realmente não pode ser fixo, pois o que está bom para você pode não ser suficiente para mim (e vice-versa).

Ainda sobre a família r, para os mais interessados, o próprio artigo de (Lakens, 2013) e outros artigos de Espirito Santo e Daniel. (2017 e 2018) esclarecem bem.

Outras medidas de TE muito usadas em estudos clínicos e epidemiológicos são o ‘risco relativo (RR)’ e a ‘razão de chance’ (*odds ratio*; OR).

São métricas que medem a associação entre a exposição a um fator de risco e a ocorrência de um desfecho.

O RR compara a probabilidade de um evento ocorrer em um grupo exposto vs. um grupo não exposto. É usado especialmente nos estudos de coorte e ensaios clínicos.

Já a OR estima a razão entre as chances de um evento ocorrer no grupo exposto vs. grupo não exposto, sendo utilizada em estudos caso-controle e transversais. Já dei exemplo em capítulo prévio, inclusive com cálculos.

Em relação a interpretação, quando o RR ou OR são iguais a 1, é porque não há diferença (não há associação) entre os grupos. Ou seja, estar ou não estar exposto ao hipotético fator de risco, não se associa com o desfecho.

Quando o RR ou OR são maiores que 1, isso indica aumento do risco ou da chance do evento (é chamado ‘fator de risco’).

Quando o RR ou OR são menores que 1, indica redução do risco ou da chance do evento (chama-se ‘fator protetor’; nesse caso, a exposição é protetora).

Um exemplo mais intuitivo: se a exposição ao cigarro gerar RR ou OR de 5 para câncer de pulmão, significa que o cigarro é um fator de risco. Interpreta-se que o hábito de fumar está associado a um risco ou chance 5 vezes maior de desenvolver câncer de pulmão em comparação com quem não fuma.

Se, por sua vez, o exercício físico gerar um RR de 0,7 para câncer de pulmão, significa que é um fator de proteção. Pode-se interpretar que quem pratica exercício tem risco 30% (ou ‘1 –

0,7 = 0,3') menor de desenvolver câncer de pulmão em comparação com quem não exercita.

A classificação da magnitude desse efeito também não é consensual. O sistema GRADE (Brasil, 2014) considera **elevada magnitude** de efeito quando o RR > 2 (fator de risco) ou < 0,5 (fator de proteção). E considera magnitude de efeito **muito grande** quando o RR > 5 (fator de risco) ou < 0,2 (fator de proteção). Uma proposta relativamente comum é esta da tabela 2.

Tabela 2. Classificação habitual para o tamanho do efeito do risco relativo (RR) e razão de chances (OR).

Classificação do Efeito	**RR ou OR** (fator de risco)	**RR ou OR** (fator de proteção)
Nulo (ou muito pequeno)	1,10	0,91
Pequeno	1,11 – 1,49	0,71 – 0,90
Moderado	1,50 – 2,99	0,51 – 0,70
Grande (ou muito grande)	> 3,00	< 0,50

Mas pode-se facilmente perceber que essa classificação é discutível. Para ilustrar, suponha que caminhar 30 min por dia esteja associado a um RR ou OR de 0,8 para mortalidade por doenças cardiovasculares. Isso significaria que essa atividade estaria associada a uma redução de 20% no risco de morte por doenças cardiovasculares.

Considerando que a maioria das pessoas é sedentária, e que caminhar é uma prática simples, sem efeitos colaterais e pode ser gratuita, uma redução de 20% no risco tem um impacto bastante considerável.

Em termos absolutos, se 1 milhão de pessoas sedentárias falecerem por doenças cardiovasculares, poderíamos estimar que 200 mil dessas mortes poderiam (no âmbito da probabilidade) ter sido evitadas apenas com a adoção desse simples hábito.

O gradiente dose-resposta também pode ser considerado aqui. Ou seja, se o estudo mostrar que a maior exposição a um dado fator aumenta ou diminui o efeito (ex: + cigarros = + câncer), isso reforça a inferência de causalidade.

Sobre a gradação de nota para esse tópico (tamanho de efeito), oriento nota entre 0 (zero) e 10. Para estabelecer essa nota, considere as tabelas de classificação apresentadas previamente, bem como sua interpretação pessoal sobre a magnitude que contemple seus interesses e do cliente/paciente.