A arte de ler um subestudo: PROMISE, falsa promessa em diabéticos

Por: Luis Claudio Lemos Correia

Este é mais um texto da série "A *Arte de Ler um Artigo Científico*". Nesta postagem abordaremos artigos construídos a partir de subanálises de grande estudos.

No texto original desta série, propusemos um momento de preparação mental antes da leitura, no qual tentamos controlar nossos vieses pessoais e ao mesmo tempo identificar situações onde nosso ceticismo deve estar mais aguçado. Esta postagem exemplifica uma frequente situação em que devemos estar muito atentos: análises secundárias de estudos originalmente negativos.

Devemos aguçar o ceticismo pois este tipo de publicação representa, na maioria das vezes, tentativa de positivar um estudo originalmente negativo. E há tantas forma de positivar (*P-hacking, statistical fishing*) que o resultado é quase garantido se for suficientemente procurado. Isto ocorre principalmente quando as análises não são predeterminadas (pois isso dá margem a um grande número de tentativas) e quando se combina várias estratégias para dar "chance ao acaso" se apresentar. Assim, publicações

secundariamente originadas de grandes estudos representam grande fonte de ilusões mediadas por erros aleatórios, sistemáticos e baixa probabilidade pré-teste de hipóteses.

Enquanto grandes estudos originais, independente da especialidade, são usualmente publicados nas grandes revistas de medicina interna (*NEJM, JAMA, Lancet, BMJ*), suas análises secundárias aparecem nas melhores revistas das especialidades. No caso de minha especialidade, cardiologia, as subanálises de grandes estudos são frequentemente publicadas no renomados *Journal of the American College of Cardiology e Circulation*, revistas com fator de impacto 17 e 19, respectivamente. Se nos parece frequente problemas de veracidade em estudos publicados nas grandes revistas de medicina, fica ainda mais evidente a tolerância com desvios de integridade científica pelas revistas *top* das especialidades. E esta tolerância é acompanhada pela comunidade de especialistas, que sofrendo do **viés de citação**espalham entusiasticamente notícias de trabalhos (pseudo) positivos.

O exemplo didático, nos presenteado esta semana, é o subestudo do PROMISE que retestou a hipótese original no subgrupo de diabéticos. O PROMISE, publicado originalmente em 2015, foi um grande (N = 10.000) ensaio clínico randomizado que comparou duas estratégias de investigação de doença coronária em pacientes sintomáticos: pesquisa anatômica (angiotomografia) *versus* métodos funcionais. O estudo não encontrou diferença de desfechos clínicos entre os dois métodos (estudo negativo).

Agora, quatro anos depois é publicada a subanálise sugerindo que em diabéticos a angiotomografia é superior a métodos funcionais.

Contextualização Clínico-Científica

O PROMISE original (*NEJM*) estudo foi motivo de postagem neste Blog, na qual eu mencionava que não havia razão para o ônus da prova de superioridade estar apenas na tomografia. Deveria ser uma hipótese científica bidirecional. E se o fosse, a conclusão adequada seria que não foi encontrada superioridade de nenhum dos métodos. Em minha interpretação, aquele estudo derrubava o mito (implausível e não baseado

em evidências) de que métodos indiretos de pesquisa de isquemia seriam melhores do que estudo da anatomia (veja post Reflexo Óculo-isquêmico). O PROMISE foi uma evidência de qualidade, capaz de colocar a estratégia anatômica, do ponto de vista de desfechos, como um método tão bom quanto os métodos "funcionais".

Em minha opinião clínica, o resultado do PROMISE alinhado à comprovadas vantagens de se conhecer anatomia (acometimento de tronco, noção de número de vasos acometidos, carga aterosclerótica, sem falar da "extrema normalidade da tomografia normal") faz deste o método não invasivo de escolha em pacientes com suspeita de doença coronária, que não tenham probabilidade alta de calcificação coronária extrema.

Emito aqui minha visão clínica dois motivos: primeiro, acho que já temos informações científicas suficientes para não precisar fabricar mais dados em prol de nenhum dos lados; segundo, proponho sempre que avaliemos nossas "evidências internas" antes de avaliar evidências externas. Neste caso, minha tendência é pela avaliação anatômica. Até mesmo porque, salvo exceções, alterações em métodos funcionais são meras consequências de obstruções anatômicas. Melhor olhar a fonte do problema, as coronárias. Por mais grosseira que esta colocação pareça aos românticos ouvidos clínicos, é um pensamento que resulta da interface entre evidências e raciocínio clínico.

Vamos à subanálise do PROMISE que, como perceberão, de promessa não tem nada. Esta nova publicação representa uma confluência de problemas com multiplicidade, viés de confusão e baixa racionalidade da hipótese testada.

Sabemos que análises de subgrupo positivas são problemáticas em estudos negativos por variadas razões que se somam: multiplicidade de testes, amostras menores que a original, implausibilidade de interações. Tudo isso aumenta a probabilidade do erro aleatório tipo I (se valer do acaso para afirmar algo falso). Mas não bastou fazer uma análise de subgrupo, estes estudo combinou a análise com uma série de problemas, listados abaixo:

1. Esta não é simplesmente uma análise de subgrupo (diabetes *versus* não diabetes). É uma análise de subgrupo de um desfecho secundário. Sabemos que análise de um desfecho secundário é outra forma de fabricar significância estatística. Pois bem, o estudo combina em uma única análise duas maneiras de fabricar o erro tipo I.
2. Para piorar, o estudo fez a análise de subgrupo não apenas de um desfecho secundário, mas de dois desfechos secundários. O primeiro deu negativo (P da interação = 0.10) e o outro deu positivo (P da interação = 0.02). Portanto, multiplicou ainda mais a multiplicidade já contida na combinação de subgrupo com desfecho secundário.
3. A análise do subgrupo "diabetes" não foi definida *a priori.* No protocolo publicado, a análise pré-especificada se referia ao desfecho “DAC equivalente” (combinado de diabetes, doença vascular periférica e doença cérebro vascular), que foi negativa no estudo primário. Portanto, essa é uma análise *post hoc* (inventada depois - nunca se sabe quantas invenções se tenta nos bastidores), o que inflaciona o problema das múltiplas comparações.
4. Finalmente, os autores criam um erro sistemático (viés): diferente do que foi feito no trabalho original, esta é **não** é uma análise por “intenção de testar” (equivalentemente à “intenção de tratar dos estudos de tratamento), a qual preveniria viés de confusão *a posteriori da* randomização. Na verdade, eles fazem uma “análises por protocolo” na medida em que excluem 1.000 pacientes que julgaram inadequados para a interpretação do exame. Exclusão de pacientes após a randomização traz risco de eliminar a homogeneidade das amostras. Principalmente quando o critério de exclusão carrega consigo pacientes de maior risco: um dos maiores motivos da dificuldade de análise da angiotomografia é a presença de calcificação coronária, a qual traduz carga aterosclerótica. Portanto no grupo tomografia provavelmente foram excluídos pacientes com maior carga aterosclerótica do que a média do grupo. Por outro lado, isso não acontece nos demais métodos não invasivos,

> nos quais as dificuldades de interpretação decorrem de problemas extra-cardíacos como janela ecocardiográfica ou atenuações mamárias/diafragmáticas. Sendo assim, a exclusão de pacientes da análise retirou indivíduos com maior carga aterosclerótica do grupo tomografia e retirou pacientes aleatórios do grupo funcional. Isso pode ter sido responsável pela pequena diferença absoluta de 1.5% observada no PROMISE. É bem verdade que esse viés pode ter ocorrido também no grupo de não diabéticos, onde não foi observada a diferença entre os métodos. No entanto, diabéticos podem ter mais calcificação do que não diabéticos, o que faria com que o viés da "análise por protocolo" fosse mais forte nos diabéticos.

Desta forma, este é um trabalho secundário com alto risco de erro aleatório e viés. Trabalhos secundários nunca devem ser vistos como confirmatórios, porém a esses pode restar um valor de sugerir uma ideia. No caso presente, fica difícil dar qualquer credibilidade ao subestudo.

Até aqui apresentei uma simples e quase óbvia discussão metodológica da validade interna deste trabalho. Porém esta discussão se torna mais interessante quando vamos além do metodológico, abordando a qualidade da ideia testada. Pois o **valor preditivo positivo de um estudo depende não só da qualidade do trabalho, como também da qualidade da ideia.**

A Qualidade da Ideia

Em primeiro lugar, devemos nos lembrar que o fenômeno de interação (*modificação de efeito*) é raro no campo científico biomédico (vejam post A Ilusão de Interação). Percebam como as análises de subgrupo de estudos positivos ou negativos quase sempre mostram consistência de resultado. Isto ocorre pois **medidas relativas de risco não sofrem modificação com o risco absoluto da população.** Para diferentes estratos de risco absoluto, a regra é observamos uma mesma redução relativa de risco, que representa a propriedade intrínseca da intervenção. Para quem continua na dúvida dessa propriedade, comparem a redução do risco relativo da estatina ou aspirina de pacientes de prevenção primária *versus*secundária. É a mesma redução relativa de risco, o que muda é o NNT, que é uma medida de

impacto individual.

Portanto, em geral uma análise de subgrupo de um estudo negativo já parte de uma baixa probabilidade pré-teste.

Partimos agora para a plausibilidade específica da ideia. Qual a diferença de diabéticos em relação a não diabéticos que poderia ter motivado estudar aquele subgrupo específico? Os autores comentam na introdução que diabéticos têm maior risco absoluto. **Aí está o erro da hipótese. Risco absoluto não promove interação!** Não há porque a tomografia ser um método igual aos demais em não diabéticos e se mostrar melhor em não diabéticos. A não ser que diabéticos tivessem menos calcificação, o que não é o caso.

Em geral, um verdadeiro efeito de interação é pouco provável em um estudo originalmente negativo. Há exceções em casos de grande plausibilidade. Neste presente caso, estamos diante da possibilidade de um diagnóstico que leva a um tratamento para doença coronariana, beneficiando o paciente no final da cascata de causalidade. Não há razão biológica para "acreditar" que a uma abordagem não teria efeito algum nos pacientes em geral, mas em diabéticos surgiria um efeito evidente. Muito mais provável o resultado encontrado ser decorrente de acaso + viés.

A maioria dos fenômenos naturais que fazem sentido a olho nu decorrem de acaso ou viés. Daí a importância de um filtro desses ruídos que se confundem com sinais verdadeiros. O filtro é o método científico, que originalmente foi criado para eliminar estes problemas, e não para criar falsas ideias (o problema da integridade científica).

Finalmente, na análise bayesiana deste trabalho, o subPROMISE testa uma hipótese de baixa probabilidade pré-teste em um desenho de estudo de baixa qualidade, o qual não consegue elevar essa probabilidade para níveis intermediários. Portanto, esse é mais um estudo de baixíssimo valor preditivo positivo.

Não precisava ...

Os Problemas nas Entrelinhas

A nossa análise de validade interna do subestudo PROMISE traduz erros metodológicos grosseiros. No entanto, **grosseiro não é sinômino de claro.** Embora falhas metodológicas grosseiras não sejam omitidas, boa parte deles não está explícito do texto do artigo, mais sim implícito de uma forma descritiva, cabendo ao leitor elaborar um pensamento do que se tratam certas descrições. Seria quase como ler nas entrelinhas.

Neste subestudo do PROMISE não está explícito que esta é uma "análise por protocolo, de subgrupo, definida *a posteriori*, de um desfecho secundário". Vejamos ponto por ponto no texto como está escrito no artigo:

Análise de Subgrupo

Para um leitor desavisado, que costuma ler apenas o título, objetivo e conclusão, este estudo facilmente passaria por um trabalho feito apenas em diabéticos.

Título: *Stress Testing Versus CT Angiography in Patients With Diabetes and Suspected Coronary Artery Disease*

Objetivo: *The purpose of this study was to assess whether a diagnostic strategy based on coronary computed tomographic angiography (CTA) is superior to functional stress testing in reducing adverse cardiovascular (CV) outcomes (CV death or myocardial infarction [MI]) among symptomatic patients with diabetes.*

Conclusão: *In diabetic patients presenting with stable chest pain, a CTA strategy resulted in fewer adverse CV outcomes than a functional testing strategy.*

Observem, nestas que são as mais importantes sentenças do artigo, que em nenhum momento há menção de que o estudo se trata da comparação do resultado do ensaio clínico entre diabéticos e não diabéticos (interação). Faz parecer que a tomografia foi comparada a métodos funcionais em uma

única população de diabéticos. Não está explícito de que esta é uma análise secundária.

Falo sobre leitores desavisados, porém eu mesmo sou um deles. Na verdade, leio poucos artigos por semana da forma ideal. Nos demais, passo o olho só para saber o que está acontecendo. Este particularmente me chamou atenção, pois eu já conhecia o PROMISE, portanto notei logo que seria um análise de subgrupo. Mas eu poderia ser enganado.

Deveria constar no título que um trabalho representa uma subanálise de um estudo original. Isso daria plena transparência. A propósito, esta é uma falha do checklist do CONSORT, que orienta apenas que o desenho geral do estudo (randomizado) seja mencionado no título.

Seria simples deixar isso transparente escrevendo algo como: Effect of diabetes in the comparison between Stress Testing Versus CT Angiography in Patients With Suspected Coronary Artery Disease: a PROMISE Substudy.

Na descrição do trabalho também não está explícito de que esta é uma análise de subgrupo, nem explícito que é não era predefinida. Vejam o momento em que percebemos que é uma análise que compara diabetes com não diabetes:

"We used contemporary data from PROMISE (Prospective Multicenter Imaging Study for Evaluation of Chest Pain), a randomized trial of diagnostic evaluation strategy in stable outpatients with symptoms suggestive of CAD. We assessed symptomatic patients with and without diabetes."

Análise de Desfecho Secundário
"The clinical outcomes of interest included time to death/MI/unstable angina hospitalization (UAH) and CV death/MI."

Apenas mencionam dois desfechos, não determinam hierarquia entre eles. Qual o primário, qual o secundário? Resposta: nenhum dos dois, pois

o desfecho primário do estudo é o combinado de quatro desfechos. Para saber isso eu precisei voltar à publicação original e relembrar qual era do desfecho primário.

Esta é uma violação da recomendação do CONSORT, que pede que desfechos sejam definidos em primários ou secundários.

Análise por protocolo
"For the present analysis, the population of patients with an interpretable testing result was used."

Precisamos a partir disso notar que foram excluídos pacientes depois de randomizados, e que estas exclusões tendem selecionar pacientes com graus diferentes de risco. Mais uma vez, não está explícito.

Plausibilidade da hipótese
O lugar de explicitar plausibilidade da ideia é na introdução do trabalho. No entanto, a introdução do estudo fala tudo menos porque o exame poderia ter um impacto diferente em diabéticos.

Discussão
Costumo sugerir a meus alunos que não percam tempo lendo a discussão. Na verdade, o tópico “discussão” tem servido mais para atenuar defeitos do que para discutir de forma transparente o valor preditivo de um estudo. Na medida em que o autor reconhece algumas falhas, gera um senso de transparência, de confiabilidade. Mas observe que a cada questão mencionada, há sempre uma frase seguinte como se dissesse “mas isso não é um grande problema”.

Menção ao defeito: itálico
Correção do defeito: negrito

"*Although our study is post hoc and is subject to the inherent limitations of this type of analysis,* ***evaluation of testing modality and outcomes in patients with diabetes was prespecified."***

Seria pior se os desfechos não fossem pré-especificados. Mas serem desfechos pré-especificados (embora secundários) não atenua o estudo ser análise de subgrupo não pré-especificada. Uma coisa é uma coisa, outra coisa é outra coisa.

"The identification of reduced risk of CV death/MI in patients with diabetes associated with CTA randomization was based on small numbers. ***The trends toward reduced risk of death/MI/UAH and CV death/MI/ UAH in patients with diabetes undergoing CTA versus functional testing reinforce the findings seen with the endpoint of CV death/MI."***

Foi erro grosseiro ter avaliado vários desfechos e não ter dito qual seria o primário. Tentou em vários para ver qual dava significativo, o que é errado. Mas ele pega essas múltiplas comparações e faz parecer que as não significativas servem de confirmação para a significativa.

"*Slight statistical differences in some of the baseline characteristics were seen in patients without diabetes who were randomized to CTA versus functional stress testing;* ***however, the absolute differences were small and likely not clinically relevant."***

Tentando mais uma vez atenuar heterogeneidades surgidas da análise de subgrupo.

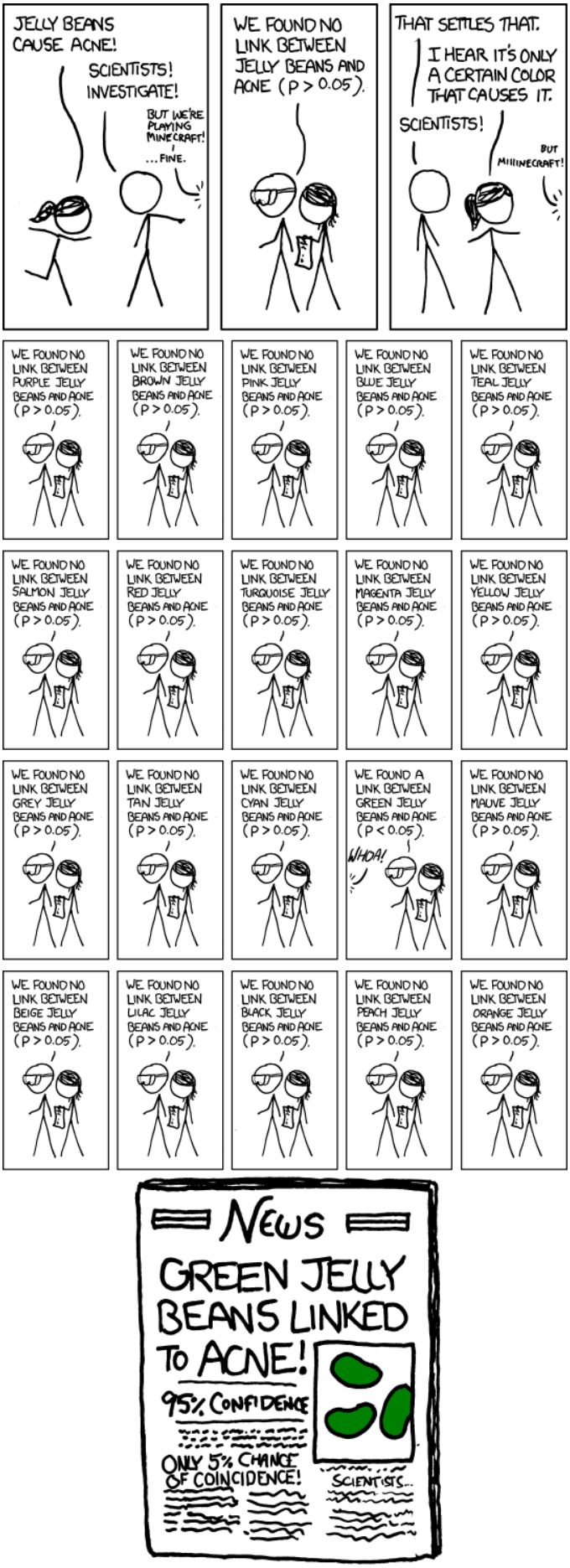
Por fim, a frase final da conclusão, que fecha tudo: *"In evaluating stable patients with diabetes who have symptoms suggestive of CAD, physicians should consider these benefits of using CTA as the initial diagnostic strategy."*

Mensagem Final

Subestudos positivos de estudos originalmente negativos nascem como uma intenção científica questionável. Devemos contrabalançar o viés da positividade (procura incessante por dados positivos) como um viés de proteção hipótese nula, a premissa científica básica.

O que piora a qualidade de Subestudos

- Baixa plausibilidade da nova hipótese testada (qualidade da ideia)
- Análises não predeterminada
- Combinação de multiplicidades (subgrupo + desfecho secundário)
- Mais de um desfecho secundário testado simultaneamente, sem hierarquia estabelecida
- Criação de vieses não contidos nos estudos originais


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...